Mechanisms of Triticeae Genome Evolution

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Romain Guyot

aus

Frankreich

Promotionskomitee

Prof. Dr. B. Keller (Vorsitz)

Prof. Dr. U. Grossniklaus

Prof. Dr. E. Martinoia

Dr. C. Feuillet

Zürich, 2004

Mechanisms of Triticeae Genome Evolution

Dissertation

zur

Erlangung der naturwissenschaftlichen Doktorwürde

(Dr. sc. nat.)

vorgelegt der

Mathematisch-naturwissenschaftlichen Fakultät

der

Universität Zürich

von

Romain Guyot

aus

Frankreich

Promotionskomitee

Prof. Dr. B. Keller (Vorsitz)

Prof. Dr. U. Grossniklaus

Prof. Dr. E. Martinoia

Dr. C. Feuillet

Zürich, 2004

Table of contents

1 Summary

In this study the main objective was to investigate mechanisms of the Triticeae genome evolution using different *in silico* approaches. This included the detailed annotation of sequenced regions in wheat and barley, as well as their comparison with sequences available in public databases to understand the colinearity relationships with other cereal genomes. The action of transposable elements on the genome evolution was studied in wheat. Two large orthologous regions on diploid and tetraploid wheat had been isolated and completely sequenced earlier. The analysis of these 427 kb of genomic sequences allowed us to discover an unusual local accumulation of Class II transposon of the CACTA type. The structure, organization and evolution of these elements as well as their impact on the wheat genome were *in silico* and experimentally investigated. Our results reveal that CACTA elements contribute significantly to the genome size and to the organization and evolution of the wheat genome. The evolution of the wheat genome was also studied by an *in silico* colinearity comparison between the rice genome and the distal part of the short arm of chromosome 1A in wheat. A total of 1.1 Mb of physical contigs from which 638 kb were completely sequenced have been generated on the wheat chromosome 1A allowing us to investigate the micro-colinearity over large distances. Many micro-rearrangements such as deletions, inversions and gene movements indicated different molecular mechanisms of grass genome evolution. In addition, *in silico* analyses revealed an intra-genomic colinearity in the rice genome suggesting an ancestral segmental duplication. The extent of such duplications has been further studied using the recent availability of the pseudo-chromosomes sequences of rice. We found that the rice genome contains extensive chromosomal duplications accounting for 53% of the available sequences. These duplications should predate the divergence of most cereals. Differential gene loss during the grass genome evolution as well as ancient large duplications in the ancestor of grass species are important factors of Triticeae genome evolution. Finally, comparative sequence analysis between β-galactosidase paralogous loci in barley revealed a mechanisms of paralogous gene inactivation such as deletion, mutation and transposon insertion in coding regions. Gene inactivation and gene loss are major factors in rapid evolutionary divergence of orthologous loci in Triticeae genomes.

2 Zusammenfassung

Das Hauptziel dieser Studie war, die Mechanismen der Genomevolution von Triticeae mit Hilfe verschiedener *in silico* Methoden zu untersuchen. Dies beinhaltete die detaillierte Annotation von sequenzierten Regionen aus Weizen und Gerste und deren Vergleich mit Sequenzen aus öffentlichen Datenbanken, um die Kolinearität mit anderen Getreidegenomen zu verstehen. Die Rolle von Transposons in der Genomevolution wurde in Weizen studiert. Zwei grosse orthologe Regionen aus diploidem und tetraploidem Weizen wurden schon früher isoliert und vollständig sequenziert. Die Analyse dieser 427 kb langen Sequenzen erlaubte es uns, eine ungewöhnliche, lokale Anhäufung von Klasse II Transposons des CACTA-Typs zu identifizieren. Die Struktur, Organisation und Evolution dieser Elemente sowie deren Einfluss auf das Weizengenom wurde *in silico* und experimentell untersucht. Unsere Resulate zeigen, dass die CACTA-Elemente bedeutend zur Genomgrösse sowie zur Organisation und Evolution des Weizengenoms beitragen. Die Evolution des Weizengenoms wurde auch mit einem *in silico* Kolinearitätsvergleich zwischen dem Reisgenom und dem distalen Teil des kurzen Armes von Chromosom 1A aus Weizen studiert. Gesamthaft wurden 1.1 Mb physisches Kontig auf dem Weizenchromosom 1A erstellt, von welchem 638 kb vollständig sequenziert wurden. Dies erlaubte uns, die Mikrokolinearität über grosse Distanzen zu untersuchen. Viele Mikro-Neuanordnungen, wie Deletionen, Inversionen und Translokationen, zeigten verschiedene molekulare Mechanismen in der Genomevolution auf. Zusätzliche *in silico* Analysen enthüllten eine intragenomische Kolinearität im Reisgenom, was auf eine frühe segmentale Duplikation schliessen lässt. Die Ausdehnung solcher Duplikationen wurde anhand der seit kurzem vorhandenen Pseudochromosomen-Sequenzen aus Reis weiter untersucht. Wir fanden, dass das Reisgenom ausgedehnte chromosomale Duplikationen enthält, welche 53% der verfügbaren Sequenzen ausmachen. Diese Genomduplikationen geschahen höchstwahrscheinlich vor der evolutiven Aufspaltung der meisten Gräser. Differentielle Genverluste während der Genomevolution der Gräser sowie frühe grosse Duplikationen in den Vorfahren der heutigen Gräserarten sind wichtige Faktoren in der Genomevolution der Triticeae. Schlussendlich offenbahrten vergleichende Sequenzanalysen zwischen paralogen Loci der β-Galaktosidase in Gerste einen Mechanismus, bei dem paraloge Gene durch Deletionen, Mutationen und Transposoninsertionen in kodierenden Regionen inaktiviert werden. Geninaktivierung und Genverluste sind zwei wichtige Faktoren in der schnellen Divergenz von orthologen Loci im Genom von Triticeae.

3 General introduction

3.1 Taxonomy and origin of the Triticeae tribe

The grass family *Poaceae* is the fourth largest family of flowering plants, comprising more than 10,000 species classified into 600 to 700 genera. The most important food crops for human nutrition and animal feed production belong to this family and include maize (*Zea mays*), sorghum (*Sorghum bicolor*), rice (*Oryza sativa*), oat (*Avena sativa*), barley (*Hordeum vulgare*) rye (*Secale cereale*) and wheat (*Triticum spp*.). Barley, rye and wheat belong to the Triticeae tribe that comprises more than 500 species in 26 genera (The NCBI taxonomy Homepage http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/). Wild progenitors of theses three cereals were initially cultivated and domesticated 12,000 years ago by pioneering farmers living in the Fertile Crescent, an area spanning, at present, Israel, Jordan, Lebanon, western Syria, southeast Turkey and along the Tigris and Euphrates rivers, into Iraq and the western flanks of Iran (Salamini et al. 2002).

Cultivated diploid barley and rye species were domesticated from their wild diploid progenitors called *Hordeum spontanum* and *Secale vavilovii,* respectively (Figure 1). In contrast, the modern bread wheat (*Triticum aestivum*) has a more complex history of domestication. *Triticum aestivum* is an allohexaploid plant carrying three different homoeologous genomes (A, B and D), that are related to a common diploid ancestor living 2.5-4.5 million years ago (MYA) (Huang et al. 2002). Bread wheat has probably been created by two independent and spontaneous hybridization events between wild species (Figure 1). The first hybridization step involved two diploid wild wheat species: *Triticum urartu* as the donor of the A genome (also called A^uA^u genome) and an extinct or undiscovered species reported to be close to the grass *Aegilops speltoides* as the donor of the B genome. Tetraploid wheat ancestors (AABB genome) appeared recently, probably 0.5 MYA to give rise to wild emmer species (*T. dicoccoides*) (Huang et al. 2002). The domestication of wild emmer was at the origin of the modern and cultivated tetraploid wheat species *T. dicoccum* and *T. turgidum* ssp. *durum.* The tetraploid wheat and the diploid wild goat grass *Aegilops tauschii,* as the donor of the D genome, were the progenitors of the hexaploid bread wheat. The hybridization between these two progenitors took probably place in the first cultivated emmer fields 8,000 years ago (Huang et al. 2002; Salamini et al. 2002).

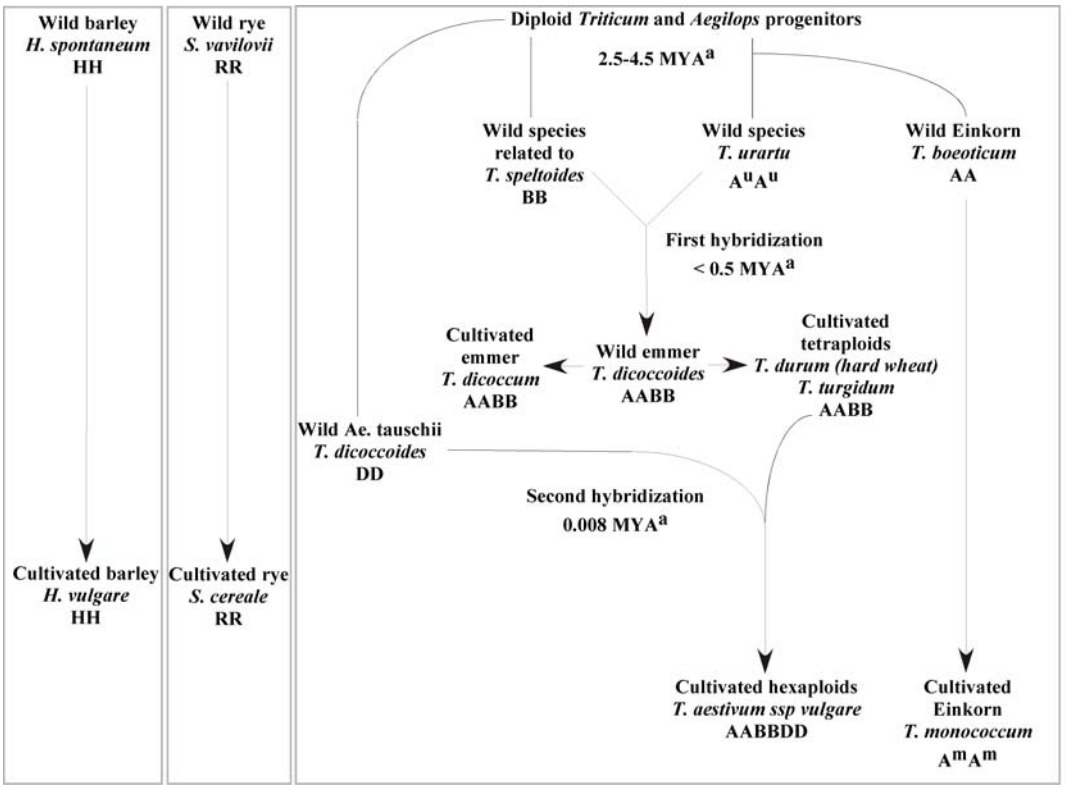

Figure 1 Genealogy of Triticeae crops
([a] Huang et al., 2002)

3.2 Genome organization in Triticeae

3.2.1 Genome size variation

Large variations of genome size exist among the different cereal species. Rice (*Oryza sativa*) has the smallest genome (~400 Mb) being three times larger than the genome of the dicotyledonous model plant Arabidopsis (Table 1). In the *Panicoideae* sub-family, sorghum and maize genomes are respectively about 2 times and 6 times larger than the rice genome. The Triticeae tribe species show a strong increase of the genome size that ranges from 4,700 Mb for *T. urartu* to more than 16,000 Mb for bread wheat (Bennett and Smith 1976) (Table 1). These genomes represent about 12 and 40 times the rice genome and about 34 and 114 times the Arabidopsis genome, respectively.

Polyploidy appears to be a very common evolutionary mechanism among flowering plants and up to 70% of all species might be polyploids (Wendel 2000). Autoploidy and allopolyploidy are the two major mechanisms that can give rise to polyploids in plants. Autopolyploidy refers to the non-disjunction of all the daughter chromosomes following meiosis in the same species. As a

consequence, the resulting progeny carries twice the initial chromosome number. Allopolyploidy refers to unreduced gametes from two distinct parental species to form a hybrid genome. The hybrids will have twice the parental chromosome number. In the Triticeae tribe, recent polyploidisations have dramatically increased the size of the bread wheat genome. The allohexaploid wheat genome (*T. aestivum*, AABBDD 2n=6x=42) was found ~25 % larger than the cultivated tetraploid wheat species *T. dicoccum* and *T. turgidum* (AABB genome, 2n=4x=28). The bread wheat genome size was also found about three times larger than the genome size of the wild or cultivated diploid progenitors (AA and DD genomes, 2n=2x=14) (Table 1).

Table 1 Comparison of the ploidy level and the estimation of genome size and repetitive DNA content among selected angiosperm species

Species	Sub-family	Monocot/ dicot	Chromosome number (2n)	Ploidy level	1C DNA amount (pg)	Estimated genome size (Mb)*	% TE
A. thaliana	*Crucifereae*	D	10	2X	0,18	174	14 [A]
Lycopersicon esculentum	*Solanaceae*	D	24	2X	1,03	994	12 [B]
Oryza sativa ssp japonica	*Oryzeae*	M	24	2X	0,4	386	14 [C, D]
Oryza sativa ssp indica	*Oryzeae*	M	24	2X	0,5	483	ND
Zea mays	*Panicoideae*	M	20	2X	2,73	2634	50-80 [E]
Sorghum bicolor	*Panicoideae*	M	20	2X	0,75	724	ND
Avena sativa	*Aveneae*	M	42	6X	13,23	12767	ND
T. urartu	*Triticeae*	M	14	2X	4,93	4757	ND
Aegilops tauschii	*Triticeae*	M	14	2X	5	4825	ND
Aegilops speltoides	*Triticeae*	M	14	2X	5,15	4970	ND
Hordeum vulgare	*Triticeae*	M	14	2X	5,55	5356	55 [F, G, H]
T. monococcum	*Triticeae*	M	14	2X	6,3	6080	ND
Secale cereale	*Triticeae*	M	14	2X	8,28	7990	ND
T. dicoccum	*Triticeae*	M	29	4X	12,03	11609	ND
T. turgidum	*Triticeae*	M	28	4X	12,28	11850	ND
T. aestivum	*Triticeae*	M	42	6X	17,33	16723	ND

From (Bennett and Smith, 1976; Bennett and Smith, 1991; Bennett and Leitch, 1995) and The Plant DNA C-value Database (http://www.rbgkew.org.uk/cval). *1pg = 965 Mb. (D) : Dicots, (M) : Monocoots, TE : Transposable Elements. (ND): not determined. A AGI, 2000; B Budiman et al., 2000; C Turcotte et al., 2001 ; D Jiang and Wessler, 2001; E SanMiguel et al., 1996 ; F Kumar and Bennetzen, 1999; G Vicient et al., 1999; H Vicient et al., 2001

Beside polyploidization, the variation of the genome size in plants is found to be largely due to varying amounts of nuclear repetitive DNA. Thirty years ago, DNA re-naturation kinetic studies indicated that repetitive DNA was the main component of large plant genomes (Smith and Flavell 1974).

In maize, repetitive DNA accounts for more than 80% of the nuclear DNA. Genomic sequencing revealed that repetitive DNA is composed of transposable elements, mainly LTR

retrotransposons (Table 1) (SanMiguel et al. 1996; SanMiguel and Bennetzen 1998). Similarly to maize, the repetitive DNA accounts for approximately 80% of the Triticeae genomes (Smith and Flavell 1974). Large-scale genomic sequencing in gene-containing regions (i.e. regions "rich" in genes and "poor" in repetitive sequences) indicated that transposable elements (TEs) account for more than 55% of the sequences, in which LTR retrotransposons represented 66.3% of all the transposable elements identified (Sabot et al. 2004). The specific amplification of LTR retrotransposons was proposed to be the main mechanism responsible for genome size expansion in Triticeae (Kumar and Bennetzen 1999).

3.2.2 Gene-rich and gene poor regions

The first data on gene distribution in Triticeae came from studies on the genome organization of hexaploid wheat (Gill et al. 1996a; Gill et al. 1996b; Sandhu et al. 2001; Sandhu and Gill 2002a). Based on the capacity of the hexaploid bread wheat genome to tolerate a certain degree of aneuploidy, more than 400 single-break wheat deletion lines have been isolated and identified. Deletion lines were used to construct a physical map for each of the 21 chromosomes of bread wheat and to identify the gene-containing regions (Endo and Gill 1996). Recently, a set of wheat deletion stocks comprising nullisomic-tetrasomic, ditelosomic and single break deletions lines has been extensively characterized at the molecular level (Qi et al. 2003). One hundred and fifty nine chromosome bins were defined among the 21 wheat chromosomes and more than 7,000 Triticeae ESTs have been physically mapped (http://wheat.pw.usda.gov/NSF/progress_mapping.html).

On each chromosome, eighty five percent of the wheat genes are clustered in six to eight different regions, representing less than 10% of the total chromosome length. This unequal gene distribution suggested the presence of gene-rich regions interspersed by large gene-poor regions (Figure 2). Similar physical locations of gene-rich regions were found among homoeologous genomes in bread wheat (Sandhu and Gill 2002a).

Deletion lines do not exist in diploid species. In barley, the genome organization was studied using the physical location of 240 translocation breakpoints integrated into the genetic map. This study demonstrated that barley chromosomes were also partitioned into gene-rich and gene-poor regions. The locations and the relative size of gene-rich regions were found similar to that of wheat suggesting a common overall genome organization in Triticeae (Kunzel 2000). This gene

organization was confirmed recently with a physical map using ultra-sensitive fluorescence *in situ* hybridization (FISH) for detecting genetically mapped cDNA clones (Stephens et al. 2004).

Comparative analysis and sequence data from gene-containing regions suggested a further sub-division into gene-rich and gene-poor compartments (Figure 2). Sequencing of large genomic sequences of gene-rich regions in wheat and barley, accounting now for more than 4,268 kb led to the identification of 164 genes (and/or pseudo-genes, Table 2) (Sabot et al. 2004). The gene density showed a mean value of 26 kb/gene which is much lower than predicted for a random genome dispersal of genes: 200-250 kb/gene (Keller and Feuillet 2000; Lagudah 2001). However, a variation of about 16 times existed among the sequenced loci. While a gene density of about 8 to 9 kb/gene similar to those observed in Arabidopsis (5 kb/gene) was found in several studies (Table 2) (Rahman et al. 1997; Feuillet et al. 2001; Brooks et al. 2002; Chantret et al. 2004), a gene density > 100 kb/gene was observed in other regions (Rostoks et al. 2002; Yan et al. 2003). This unequal gene distribution indicates that gene-rich regions are probably made up of single genes and high-gene density islands for which the gene density can locally reach up to 4-5 kb (Feuillet and Keller 1999). Despite this variation, a mean gene density of 26 kb/gene supports a non random distribution of genes in the Triticeae genomes and reinforces the hypothesis of the gene-rich/gene-poor regions structure along the Triticeae chromosomes. In gene-rich regions, single gene and gene-dense islands were found interspersed by large stretches of nested retrotransposons (Shirasu et al. 2000; Dubcovsky et al. 2001; Wicker et al. 2001; Wicker et al. 2003b). The variation of the gene density seems to be mainly due to local expansion or contraction of the transposable element content. Gene-poor regions that separate gene-rich regions in the Triticeae chromosomes, are supposed to be composed of large blocks of repetitive DNA, mainly made up of retrotransposon-like sequences (Feuillet and Keller 1999).

However, despite the fact that gene-poor regions constitute the main part of the Triticeae genome, the fine composition, structure and evolution of such regions remain very poorly understood. These regions should be targeted to large-scale sequencing and analyses to provide a better overview of the overall genome organization.

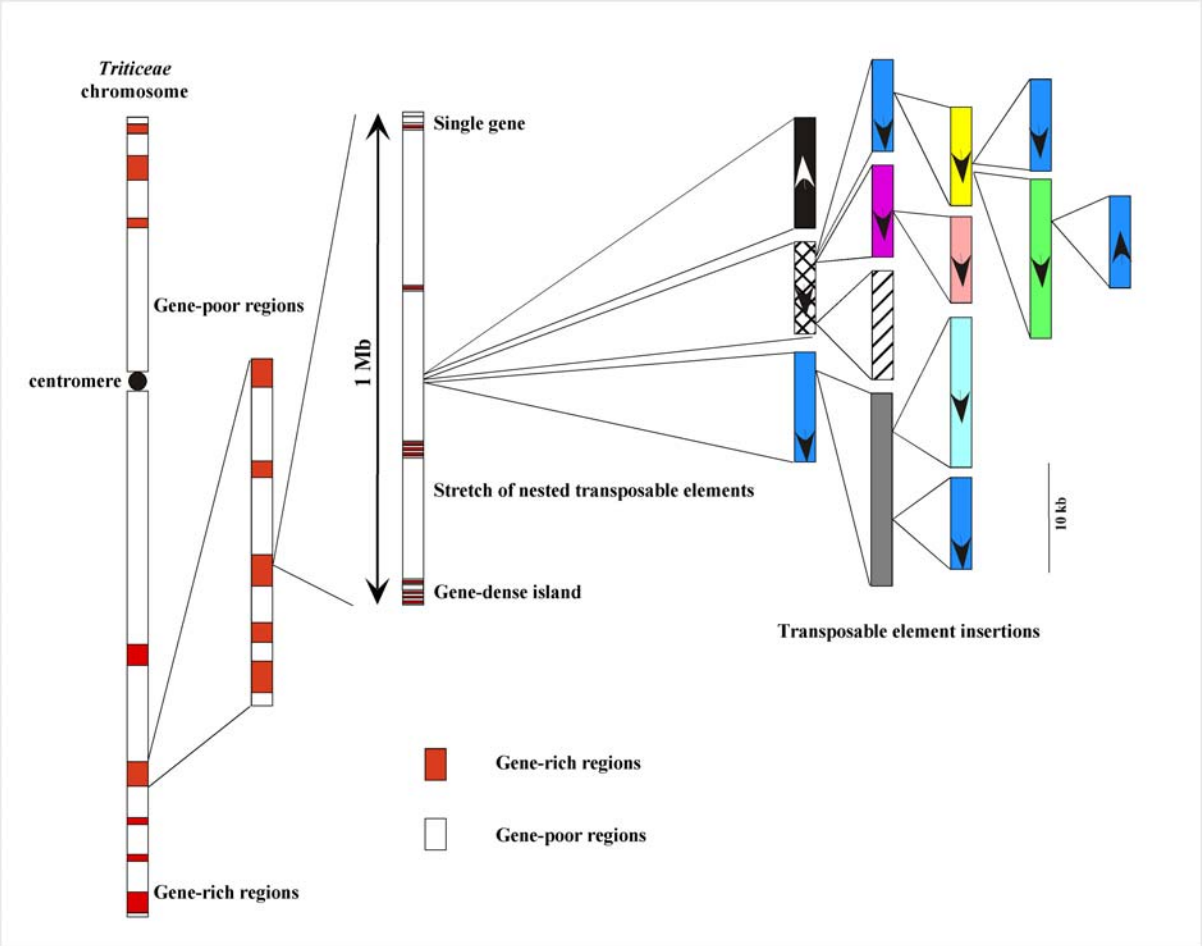

Figure 2 Model of genome organization in Triticeae

3.2.3 Transposable elements

Triticeae genomes comprise more than 80% of repetitive DNA (Table 1) and genomic sequencing in gene-containing regions has revealed that transposable elements (TEs) and derivatives are the main components of such genomes (Panstruga et al. 1998; Shirasu et al. 2000; Dubcovsky et al. 2001; Feuillet et al. 2001; Wicker et al. 2001; Brooks et al. 2002; Rostoks et al. 2002; SanMiguel et al. 2002; Wei et al. 2002; Yan et al. 2002; Brunner et al. 2003; Wicker et al. 2003b; Yan et al. 2003; Chantret et al. 2004; Gu et al. 2004; Kong et al. 2004; Yan et al. 2004). Transposable elements share several key properties such as the ability to move from one chromosome location to another or the ability to amplify their copy number within the host genome. In addition, all TEs also share a common structural feature. The integration of TEs generates a duplication of a short nucleotide sequence at the insertion site that flanks both ends of the TE. Lengths and nucleotide compositions of these sequences, called target site duplications

(TSD) are specific for a TE class and family and make them a useful tool for identification and classification of TEs (Figure 3). During a long time, TEs were often considered as selfish, parasitic and even as the "junk" DNA of plant genomes. The discovery of their rich structural and behavioral diversities, their high redundancies and their important contribution to the genome organization indicates their central role in the dynamic evolution of the Triticeae. Initially, TEs were classified into two major classes based on their mode of replication. Each class was found composed of both autonomous and non-autonomous members. Class I elements or retrotransposons move through an RNA intermediate whereas class II elements or transposons move directly through a DNA form by a "cut and paste" mechanism.

Recently, the isolation and sequencing of large genomic regions allowed to discover new types of TEs and, based on these new structures, the initial classification was refined. More new TEs are again expected to be discovered with the increasing availability of sequences from large plant genomes such as Triticeae genomes.

3.2.3.1 Class I transposable elements

Class I transposable elements or retrotransposons group together TEs that move via an RNA molecule which is reverse-transcribed into an extra-chromosomal DNA and inserted elsewhere in the genome (Kumar and Bennetzen 1999; Bennetzen 2000). This replicative mode of transposition as well as the element structure is similar to retroviral genomes. While retroviruses are found only in animals, retrotransposons are the most abundant class of TEs in all eukaryotic genomes. Retrotransposons make-up 50 to 70 % of the nuclear genome of maize (SanMiguel and Bennetzen 1998) and an even higher proportion is expected in the Triticeae genomes considering that retrotransposons from the main part of TEs in gene-rich regions of wheat and barley (Sabot et al. 2004). Their mode of replication that creates copies of elements can explain such predominance in plant genomes.

Class I TE is divided into LTR retrotransposons and non-LTR retrotransposons (Figure 4). LTR retrotransposons are further sub-classified into Ty1-*copia* and Ty3-*gypsy* according to the order of the encoded integrase (int) and reverse transcriptase (RT) / RNAse H gene products. In each sub-class, elements are again sub-classified depending of their sequence similarities. Recently, two new types of LTR retrotransposons were identified: the TRIM (Terminal-repeat

retrotransposons in miniature) (Witte et al. 2001) and the LARDS elements (Large retrotransposon derivatives) (Kalendar et al. 2004).

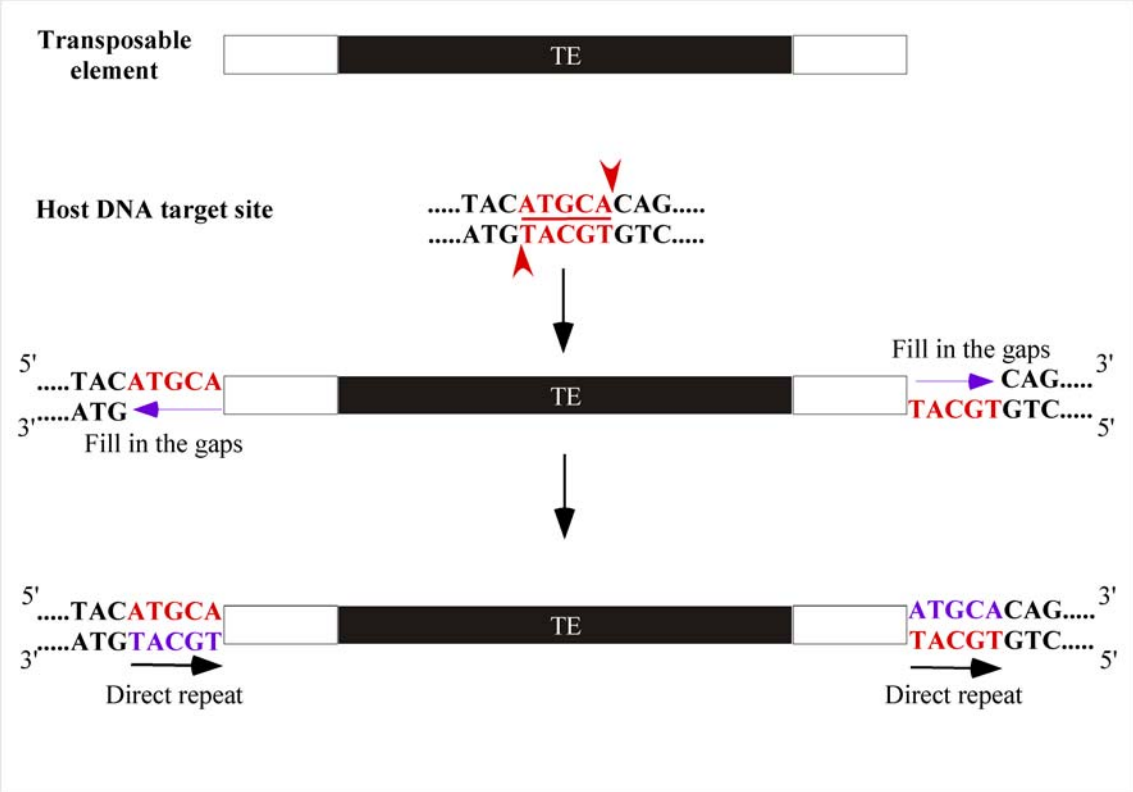

Figure 3 Generation of Target Site Duplications (TSD) upon the insertion of transposable elements

LTRs (Long Terminal Repeat) of LTR retrotransposons are direct repeats, flanking the coding regions of the retrotransposon and ranging in size from several hundred base pairs to more than five kb. Each LTR contains promoter and poly-adenylation sequences that are essential for the transcription and reverse transcription of the element. LTRs terminate with a short inverted repeat, usually 5'-TG(N)-3' and 5'-(N)CA-3'. Together with a typical TSD length of 5 bp, they constitute important features for the identification and the classification of elements.

Proteins are synthesized as polyproteins that are cleaved into functional peptides. The *pol* gene encodes the reverse transcriptase (RT), the RNAse-H proteins required during the replication of the transposon and the integrase protein (int) involved in the integration of the retrotransposon in a new chromosomal location in the genome.

Table 2 Gene densities observed in wheat and barley large genomic sequences

Species Group	Species	Locus	Genome Group	Accession	Sequence length (kbp)	Genes and pseudo-genes	Gene density kb/gene	References
Wheat	*Ae.tauschii*	*HMW-Glu*	DD	AF497474	102,8	4	25,7	(Anderson et al., 2003)
Wheat	*Ae.tauschii*	*LZ-NBS-LRR*	DD	AF446141	106,6	11	9,7	(Brooks et al., 2002)
Wheat	*Ae.tauschii*	*LRR*	DD	AF532104	27,9	2	14	(Huang et al., 2003)
Wheat	*Ae.tauschii*	*SBE*	DD	AF525764	25,1	3	8,4	(Rahman et al., 1997)
Wheat	*Ae.tauschii*	*3DS LRR*	DD	AY534122	26	3	8,7	Whitford et al, unpublished
Wheat	*Ae.tauschii*	*3DS LRR*	DD	AY534123	198,6	7	28,4	Whitford et al, unpublished
Wheat	*T.monococcum*	*215kb-5A*	A^mA^m	AF459639	215,2	5	43	(SanMiguel et al., 2002)
Wheat	*T.monococcum*	*LMW-Glu*	A^mA^m	AY146588	285,4	4	71,4	(Wicker et al., 2003b)
Wheat	*T.monococcum*	*LR10*	A^mA^m	AF326781	211	6	35,2	(Wicker et al., 2001)
Wheat	*T.monococcum*	*Vrn1*	A^mA^m	AY188331	133,6	1	133,6	(Yan et al., 2003)
Wheat	*T.monococcum*	*Vrn1*	A^mA^m	AY188332	95,5	3	31,8	(Yan et al., 2003)
Wheat	*T.monococcum*	*Vrn1*	A^mA^m	AY188333	112,3	1	112,3	(Yan et al., 2003)
Wheat	*T.monococcum*	*Vrn2*	A^mA^m	AY485644	438,8	10	43,9	(Yan et al., 2002; Yan et al., 2004)
Wheat	*T.monococcum*	*Ha*	A^mA^m	AY491681	101,1	11	9,2	(Chantret et al., 2004)
Wheat	*T. durum*	*LMW-Glu*	AA	AY146587	142	3	47,3	(Wicker et al., 2003b)
Wheat	*T. durum*	*HMW-Glu*	AA	AY494981	307	4	76,8	(Gu et al., 2004)
Wheat	*T. durum*	*HMW-Glu*	BB	AY368673	285,5	7	40,8	(Kong et al., 2004)
Wheat	*T. aestivum*	*LRR*	BB	AF325196	35,8	4	9	(Feuillet et al., 2001)
Wheat	*T. aestivum*	*LRR*	BB	AF325197	20,7	2	10,4	(Feuillet et al., 2001)
Wheat	*T. aestivum*	*LRR*	DD	AF325198	43,6	4	10,9	(Feuillet et al., 2001)
Barley	*H. vulgare*	*Rar1*	HH	AF254799	65,9	3	22	(Shirasu et al., 2000)
Barley	*H. vulgare*	*Mla*	HH	AF427791	261,2	17	15,4	(Wei et al., 2002)
Barley	*H. vulgare*	*745c13*	HH	AF474071	102,8	1	102,8	(Rostoks et al., 2002)
Barley	*H. vulgare*	*773k14*	HH	AF474072	113,5	4	28,4	(Rostoks et al., 2002)
Barley	*H. vulgare*	*259I16*	HH	AF474373	124	10	12,4	(Rostoks et al., 2002)
Barley	*H. vulgare*	*11o09*	HH	AF474982	77,1	5	15,4	(Rostoks et al., 2002)
Barley	*H. vulgare*	*Rph7*	HH	AF521177	211,6	10	21,2	(Brunner et al., 2003)
Barley	*H. vulgare*	*635P2*	HH	AY013246	102,4	5	20,5	(Dubcovsky et al., 2001)
Barley	*H. vulgare*	*HMW-Glu*	HH	AY268139	120,5	4	30,1	(Gu et al., 2004)
Barley	*H. vulgare*	*Vrn2*	HH	AY485643	114,9	7	16,4	(Yan et al., 2002)
Barley	*H. vulgare*	*Mlo*	HH	Y14573	59,7	3	19,9	(Panstruga et al., 1998)
Total					**4268**	**164**	**26**	

The most abundant TEs in the Triticeae genome are members of the *BARE-1* (Barley RetroElement 1) super-family, which belongs to the LTR retrotransposons Ty1-*copia* group. It was first described as a transcriptionally active element in barley (Manninen and Schulman 1993; Vicient et al. 1999; Vicient et al. 2001) and was named *Wis-2* (Harbert 1987; Murphy et al. 1992) or *Angela* (Wicker et al. 2001) in wheat. *BARE-1* has a size of about 12 kb and more than 10,000 copies were estimated to be present in barley, contributing to 3% of the whole genome (Vicient et al. 1999). In contrast to active *BARE-1* elements, the main part of the identified LTR

retrotransposons from Triticeae genomic sequences was found inactive. Truncated, fragmented and degenerated elements composed the majority of all the elements identified so far. In addition, complete LTR retrotransposons were often found inactivated by accumulation of mutations responsible for stop codons in open reading frames as well as for frame shift mutations.

TRIM is a new group of TEs that are highly redundant and ubiquitous in the plant kingdom (Witte et al. 2001). These small elements (~500 bp) show common structural features of LTR retrotransposons such as the presence of two LTRs and internal domains containing recognition motifs required for their mobility (Figure 4). However, TRIMs are non-autonomous elements because they lack coding capacities and require proteins encoded in *trans* for their transposition. An unequal recombination event from two identical LTR retrotransposons was proposed to be at the origin of such elements (Witte et al., 2001).

LARDS form a distinct group of LTR retrotransposons (Kalendar et al. 2004). They are structurally similar to TRIM with the presence of LTR and their lack of coding sequences, but they are unique because they carry long and conserved non-coding internal sequences. LARDS are non-autonomous elements as well, which can be activated by proteins encoded elsewhere in the genome by complete and autonomous elements of the same family. LARDS possibly originated from Ty3-*gypsy* LTR retrotransposons by unequal recombination mechanisms (Kalendar et al. 2004).

Non-LTR retrotransposons group together LINE and SINE elements standing for respectively Long and Short Interspersed Nuclear Elements (Kumar and Bennetzen 1999; Schmidt 1999). They are also ubiquitous components of plant nuclear genomes but probably less active than LTR retrotransposons.

LINEs are several kb long elements and contain two open reading frames, encoding enzymes necessary for their autonomous transposition. The absence of LTRs and integrase domain as well as the presence of a polyA tail and untranslated regions (UTR) at both ends clearly separate LINEs from LTR retrotransposons (Figure 4). It has been proposed that LINEs are the oldest class of retrotransposons and that LTR retrotransposons may have arisen by the acquisition of the LTR structure by LINEs (Kumar and Bennetzen 1999).

SINE is the second type of non-LTR retrotransposons. SINEs are short (100 to 300 bp) compared to LINEs and they carry structures similar to those of tRNA called boxes A and B. These boxes serve as an internal promoter for the recognition of the polymerase III transcription complex.

Based on these structures, it has been proposed that SINE derive from polymerase III genes such as tRNA (Schmidt 1999).

SINEs do not encode their own reverse transcriptase and they are unable to transpose autonomously. They require enzymes encoded in *trans* by autonomous LINEs or LTR retrotransposons for their mobility. So far, few SINEs have been identified in plant genomes.

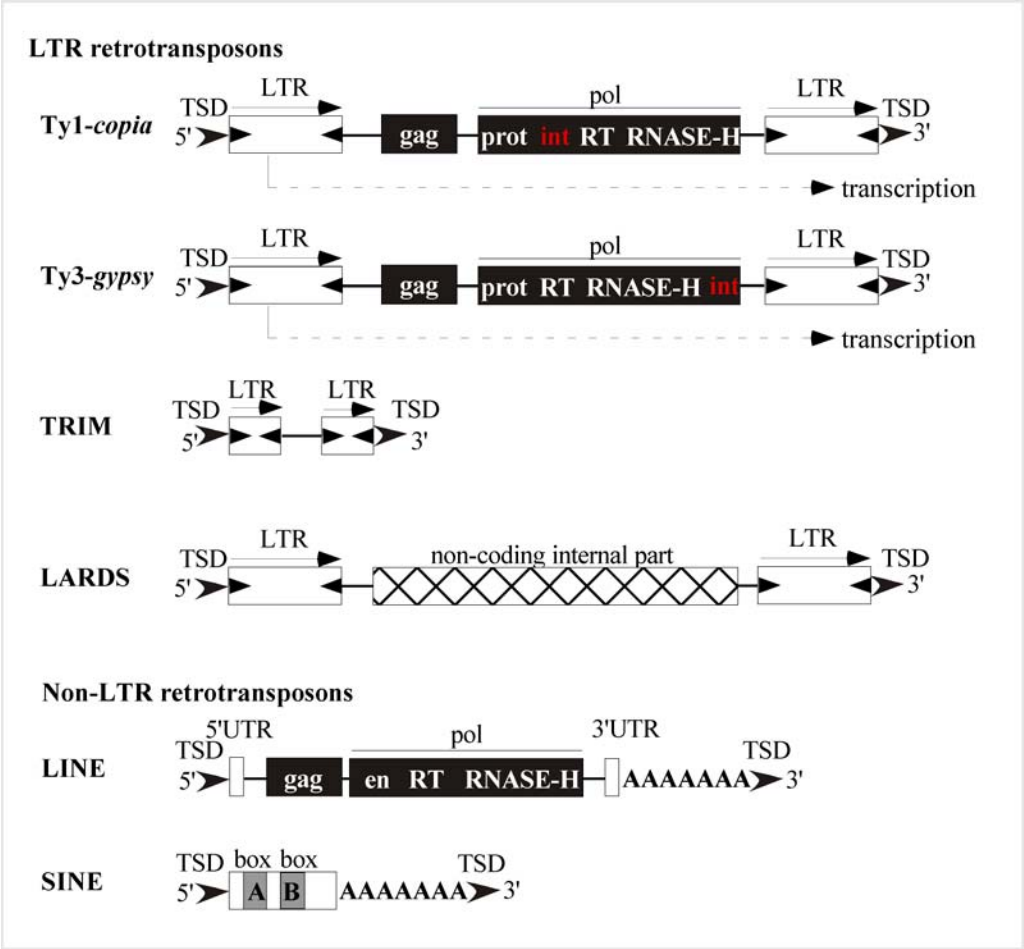

Figure 4 Structure of the major types of class I plant transposable elements
Genes of retrotransposons encode capsid-like proteins (*gag*), endonuclease (*en*), protease (*prot*), integrase (*int*), reverse transcriptase (*RT*) and *RNAse-H*.

3.2.3.2 Class II transposable elements

Class II transposable elements move via a DNA intermediate (Le et al., 2001; Feschotte et al., 2002). Transposons, MITES (Miniature inverted-repeat transposable elements) and Foldback elements are grouped together into Class II. The class II transposons group includes the *hAT* (*hobo*, *Activator* and *Tam3* (Rubin et al. 2001), CACTA (Nacken 1991), *Mutator*, PIF/Harbinger

(Zhang et al. 2001) and TC1/mariner (Feschotte and Wessler 2002) superfamilies (Table 3). The first transposon element, belonging now to the *hAT* superfamily, was identified in a site of chromosome breakage in maize and is called *Ds* (Dissociation). *Ds* can transpose or break chromosomes only in the presence of another locus called *Ac* (Activator).

Together *Ac* and *Ds* constitute a TEs family that is representative of the diverse structure of class II elements composed of autonomous (*Ac*) and non-autonomous elements (*Ds*) (Table 3). Non-autonomous elements lack coding regions and depend on the machinery of proteins encoded elsewhere in the genome for their transposition. Transposons are found in all organisms and most of the elements range in size from several hundred bp to more than 10 kb. They carry Terminal Inverted Repeats (TIRs) at both ends of the element that range in size from 11 bp (*Ac/Ds*) to several hundred bases (Figure 5).

The complex pattern of sub-terminal tandem and inverted repeats of TIRs is the specific target site for the recognition of transposase enzymes involved in the excision and the integration mechanisms. Transposons can encode a single or multiple genes that catalyze and regulate their transposition. A family of transposons is defined and classified by the structure, sequence and size of their TIRs, the length of TSD and the sequence conservation of the transposase gene encoded by autonomous elements.

MITEs form a group of abundant, non-autonomous elements. Despite a structure similar to transposons (Figure 5), their small size (<500 bp), high copy number in gene-rich regions of plant genomes and their specific TSD length and composition (usually 2 or 3 bp: TA or TAA) allow to clearly distinguish them from other class II transposable elements. Based on the sequence conservation of their TIRs, MITEs are mainly classified into two large families: *Stowaway-like* and *Tourist-like* (Bureau and Wessler 1994a; Bureau and Wessler 1994b; Feschotte et al. 2002). Evidence for insertion and excision of a MITE came recently from the rice genome (Jiang et al. 2003; Kikuchi et al. 2003a; Nakazaki et al. 2003). The MITE called *mPing* (*PIF-like*/*Tourist-like* family, Table 3) is a deletion derivative from an autonomous transposon (*Pong*) and is able to use proteins encoded elsewhere for its activity. These results indicate that MITEs can derive from transposons.

Foldback are non-autonomous transposable elements that share common structures with non-autonomous transposons and MITEs.

Table 3 Transposon super-families in plants

Transposon superfamily	Species	Autonomous members	Non-autonomous members	Copy number	References
hAT	*Zea mays*	*Ac*	*Ds*	50-100	(Wessler, 1988)
CACTA	*Zea mays*	*Spm*	*dSpm*	50-100	(Gierl, 1996)
	A. thaliana	*CAC1*	*CAC2*	4	(Miura et al., 2001)
Mutator	*Zea mays*	*MuDR*	*Mu1*	10-100	(Chandler et al., 1986)
	A. thaliana	*AtMu1*	-	1	(Singer et al., 2001)
PIF/Habinger	*Zea mays*	*PIFa*	*mPIF*	6000	(Zhang et al., 2001)
	Angiosperm	*PIF-like*	*Tourist-like*	Variable	(Zhang et al., 2001)
	O. sativa	*Pong*	*mPing*	14-70	(Jiang et al., 2003)
Tc1/Mariner	*Angiosperm*	*MLEs*	*Stowaway-like*	Variable	(Turcotte, 2001; Feschotte and Wessler, 2002)

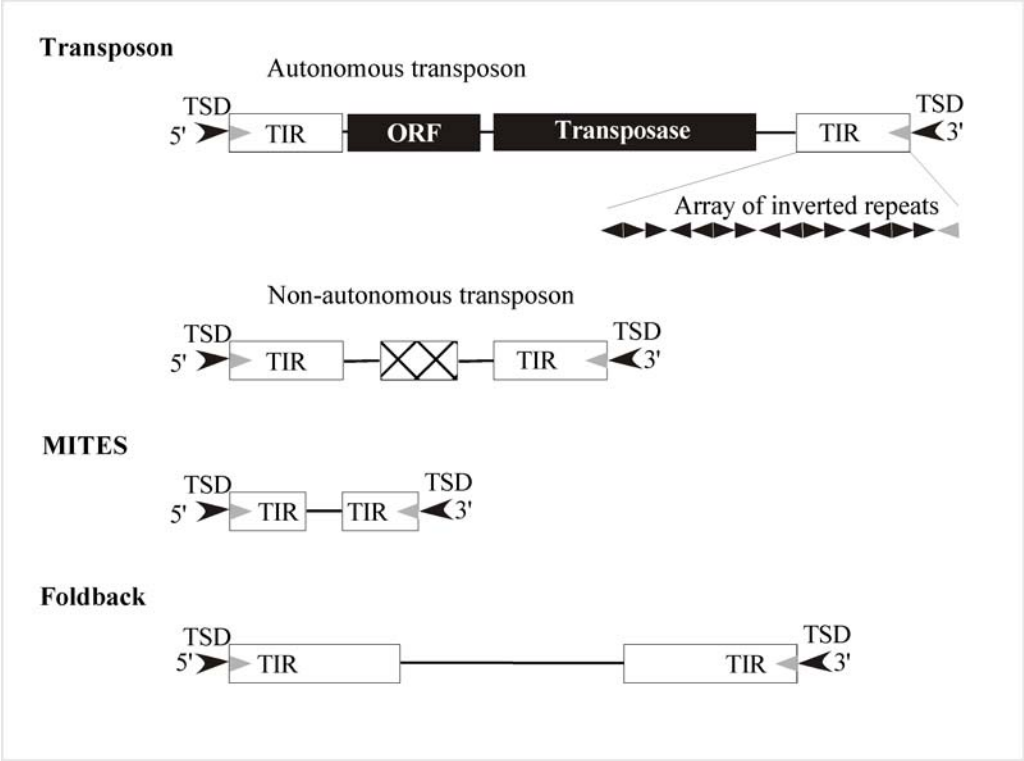

Figure 5 Structure of the major types of class II plant transposable elements

An internal and usually AT-rich region is surrounded by long TIRs of several hundred bp length. TIRs, consisting of tandemly arranged repeats are flanked by TSD with variable composition and length (Rebatchouk and Narita 1997). Little is known about the foldback elements in plant genomes. They are believed to be ancestral transposon derivatives which have lost coding capacities.

3.2.3.3 Unclassified elements

Unclassified TEs show neither coding capacities nor common structural features of TEs classified previously. These elements were identified by the presence of multiple and complete or partial copies in available sequence databases and, in many cases, the presence of a TSD is often the only indication that they represent mobile DNA elements. In some cases, the identification of the disruption of characterized nucleotide structures by an unknown sequence flanked by a TSD led to the identification of these unclassified elements. Some of them could be remnants of deletion or recombination that have probably diverged to the point of being unrecognizable as members of transposable elements. Large-scale genome sequencing will allow to identify the full-length or complete elements and then to discover their overall structures.

3.2.4 Organization of TEs in plant genomes

Knowledge about the chromosomal distribution of TEs is important for understanding plants chromosome structure and organization as well as the role of TEs in evolutionary processes. The mobility, the insertion site preferences as well as the high structural diversities have led to specific arrangements of transposable elements along plant chromosomes. Different plant genomes with different sets and numbers of TEs families can have different composition and arrangements of TEs. While small genomes such as Arabidopsis have very few TEs that are mainly clustered in centromeric and telomeric regions (AGI. 2000), larger plant genomes such as maize, wheat and barley have mostly TEs interspersed along chromosomes.

Retrotransposons were found clustered in heterochromatin of Triticeae chromosomes (Figure 6). While Ty3-*gypsy* LTR retrotransposons are one of the main component of centromeres (Presting et al. 1998; Kumar and Bennetzen 1999; Hudakova et al. 2001) Ty1-*copia* LTR retrotransposons and LINEs were localized in the terminal heterochromatin of wheat and barley chromosomes (Belyayev et al. 2001). In addition, a combination of both Ty1-*copia* and Ty3-*gypsy* was also observed in the pericentromeric heterochromatin in Triticeae (Belyayev et al. 2001).

Retrotransposons were also found dispersed along euchromatic regions of chromosome arms. Recent sequence data in Triticeae gene-rich regions have revealed more on the local organization of TEs in such regions. Retrotransposons, mainly LTR retrotransposons, were found located in

intergenic regions and organized in clusters or long stretches of elements inserted within each others in a complex patchwork. This pattern can be explained by the fact that LTR retrotransposons act as preferential target sites for the insertion of other LTR retrotransposons (SanMiguel et al. 1996; Tikhonov et al. 1999; Vicient et al. 1999; Shirasu et al. 2000).

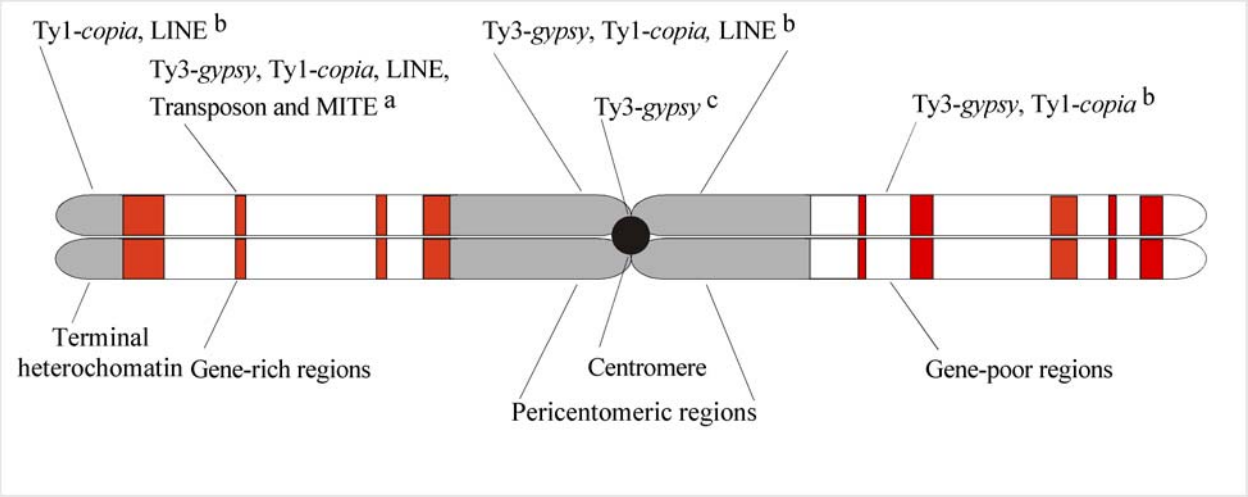

Figure 6 Distribution of TE in a Triticeae model chromosome
[a] Brooks et al., 2002; Brunner et al., 2003; Chantret et al., 2004; Dubcovsky et al., 2001; Feuillet et al., 2001; Gu et al., 2004; Kong et al., 2004; Panstruga et al., 1998 ; Rostock et al., 2002; SanMiguel et al., 2003; Shirasu et al., 2000; Wei et al., 2002; Wicker et al., 2001 ; 2003; Yan et al., 2002 ; 2003 ; 2004
[b] Belyayev et al., 2001
[c] Presting et al., 1998; Hudakova et al., 2001; Ito et al., 2004

Class II TEs exhibit a preferential insertion within euchromatic regions, close to the gene space. In maize, the *Mutator* transposons target specifically the genic sequences as well as low-copy number DNA (Cresse et al. 1995). In contrast, very few large transposons have been identified in Triticeae genomes so far and their distribution remains speculative. However, numerous MITEs have been identified and they are clearly associated with coding regions similarly to barley, maize, rice and sorghum genomes (Bureau and Wessler 1994a; Bureau and Wessler 1994b; Iwamoto and Higo 2003; Sabot et al. 2004).

3.2.5 Contribution of TEs to the evolution of plant genes and genomes

TEs are potent mutagenic agents and can induce a wide range of changes in host genomes. Mutations induced by TEs range from single nucleotide insertion or deletion to large rearrangements such as deletions, duplications and inversions. Beside these dramatic changes, TEs can be considered as genomic resources for natural selection and biodiversity.

3.2.5.1 Gene mutation

Class I TEs can induce gene mutations through their insertion near or within genes. In wheat, a High Molecular Weight Glutenin gene (*HMW-Glu-1*) located on chromosome 1A was found silenced by the insertion of a Ty1-*copia* LTR retrotransposon *WIS 2-1A* (Harbert 1987). Active retrotransposons can also affect transcription of adjacent genes in a negative manner by producing sense or anti-sense transcripts of those genes (Kashkush et al. 2003). However, the clustering of LTR retrotransposons in long stretches of nested elements could restrict the deleterious effect of their activity on genes. Class II TEs are well known to create phenotypic mutations upon their insertion. They were first discovered by Barbara McClintock as responsible for the sector of altered pigmentation on maize mutant kernel (http://profiles.nlm.nih.gov/LL/Views/Exhibit/visuals/origins.html). In addition, transposons carry their own regulatory sequences that can also alter expression of neighboring genes if they are inserted within the promoter of the considered gene (Bennetzen 2000).

3.2.5.2 Genome size

TEs play a major role in the size of plant genomes. While small genomes such as the one of Arabidopsis have a low content of TEs, the large genomes from maize and Triticeae might be the result of a successful LTR retrotransposon proliferation. In all plants, the expansion of LTR retrotransposon copy number appears to be relatively recent. The estimation of the insertion date of LTR retrotransposons and comparative analysis at orthologous loci between maize and sorghum indicated that amplification have occurred 2 to 6 MYA in maize, i.e. after the maize and sorghum split 15 to 20 MYA (SanMiguel and Bennetzen 1998; SanMiguel et al. 1998). Similar estimations were obtained in rice, wheat and barley species (SanMiguel et al. 2002; Wicker et al. 2003b; Ma et al. 2004).

LTR retrotransposons are also involved in mechanisms of genome contraction such as unequal and illegitimate recombination that counteract genome expansion caused by retrotransposons (Figure 7). Unequal intra-strand recombination between LTRs of the same element can generate SoloLTRs (Figure 7). Unequal intrastrand recombination between two different LTR

retrotransposons in direct orientation can give a patchwork of retrotransposon structures and can also generate inter-element SoloLTRs (Figure 7).

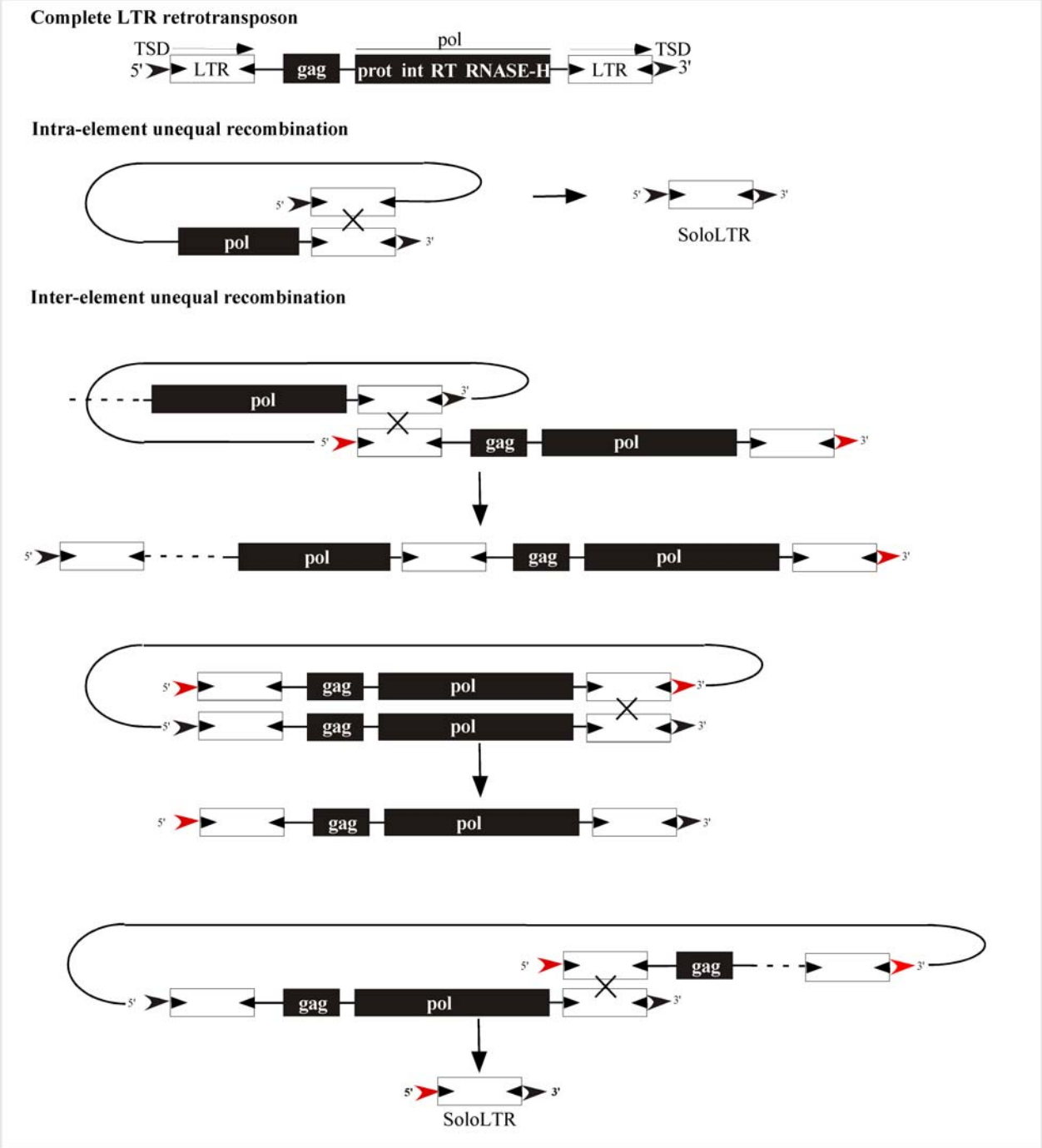

Figure 7 Unequal intrastrand recombination mechanisms between LTR retrotransposons
Unequal intra and inter-element recombination can create SoloLTRs and deletions.

These mechanisms can induce an important loss of DNA that can counteract the genome expansion. In rice, 190 Mb of DNA were estimated to be deleted in the last 8 million years by

such mechanisms (Ma et al. 2004). Frequent deletions within LTR retrotransposons that do not belong to unequal recombination mechanisms have been observed in Arabidopsis, rice and Triticeae genomes (Wicker et al. 2001; Devos et al. 2002; Rostoks et al. 2002; Wicker et al. 2003b; Ma et al. 2004). These deletions are associated with short tandem repeats at both ends of breaks suggesting that illegitimate recombination mechanisms also contribute to the reduction of genome size.

3.2.5.3 Genome rearrangements

TEs promote various chromosomal rearrangements by recombination mechanisms. Due to their high copy number and dispersal, they play a central role in the re-organization of the host genomes. Intra-chromosome homologous recombination mechanisms between elements of the same family can lead to duplication, deletion, inversion and acentric / dicentric chromosomes (Figure 8), whereas inter-chromosome recombination mechanisms cause reciprocal translocation.

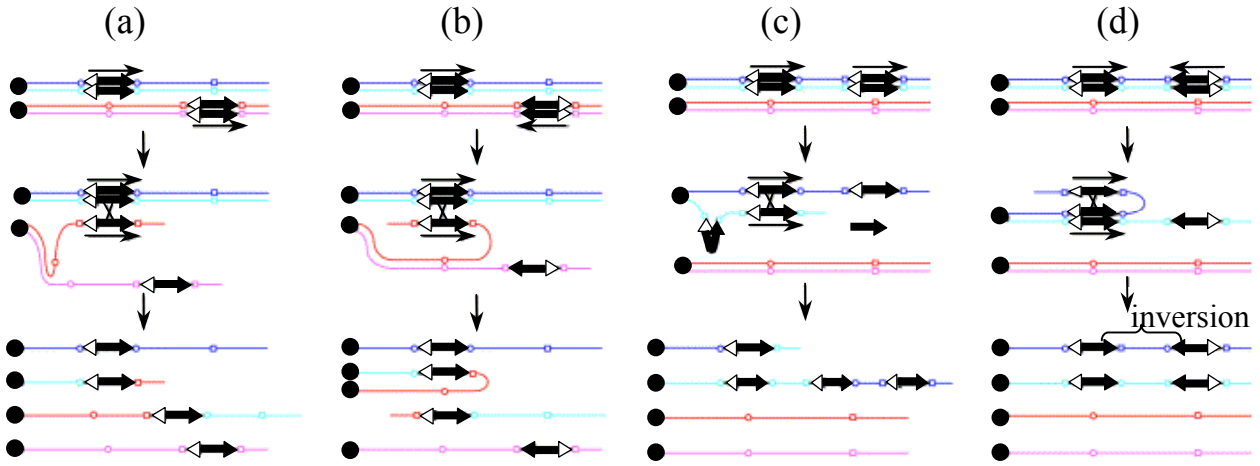

Figure 8 Chromosomal rearrangements caused by homologous recombination
Homologous recombination between repetitive sequences, such as TEs, can result in chromosomal rearrangements such as deletions, duplications and inversions. (a) Recombination between TEs in the same relative orientation on homologous chromosomes results in the formation of chromosomes containing either a deletion or a duplication of the intervening sequence. Both rearrangements are associated with recombination between two homologues. (b) Recombination between TEs in opposite relative orientation on homologous chromosomes results in the formation of a dicentric chromosome and an acentric fragment. (c) Recombination between TEs in the same relative orientation on one chromosome results in the formation of chromosomes containing either a deletion or a duplication of the intervening sequence, differing from events in (a) by the lack of recombination between homologues and the net increase or decrease of the TE number. (d) Recombination between TEs in opposite relative orientation on one chromosome can result in the formation of an inversion between the two TEs. If caused by homologous recombination, deletions and duplications can only be formed by TEs in the same relative orientation and inversions can only be formed by TEs in opposite relative orientation. From Gray, 2000.

3.3 Comparative genomics. A tool to understand the evolution of genome organization in the grass family

Comparative genomics is the comparison of genome structure and function among different genomes and the search for similarities. Comparative genomic research has the following main goals: (i) the comparison of genome structure can reveal mechanisms and processes of genome evolution and (ii) the comparison creates a genome reference for transfer of knowledge to related species.

3.3.1 Phylogeny and timescale of evolution in the grass family

To better understand the evolution of genome organization in the grass family, phylogeny and divergence times have been determined between distantly related key species.

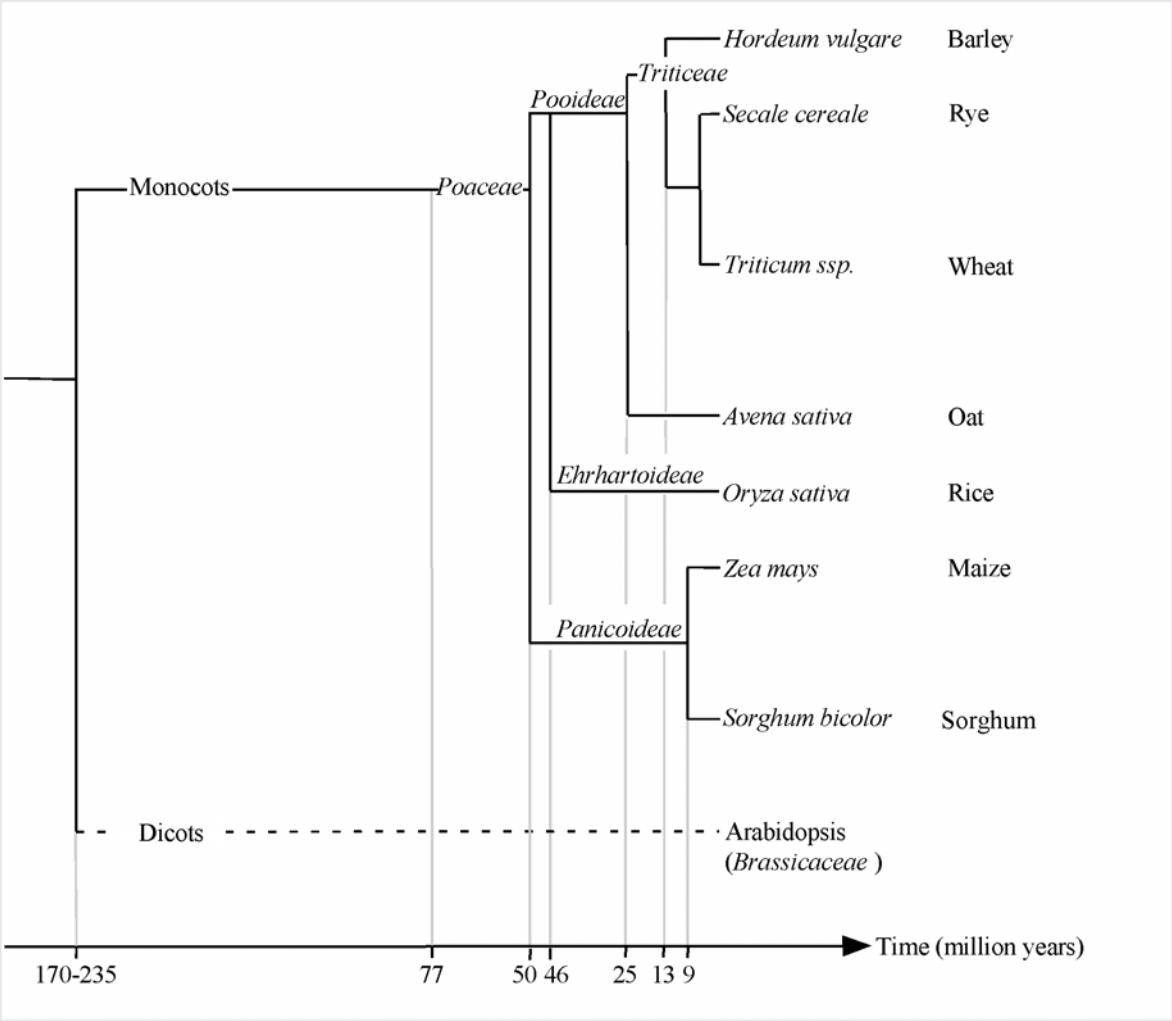

Figure 9 Phylogeny and timescale evolution of the grass family
Wolfe et al., 1989; Kellogg EA, 2001; Gaut, 2002.

The origin of grasses was estimated roughly at 77 MYA, postdating the separation between monocotyledonous and dicotyledonous plant species (Wolfe et al. 1989; Kellogg 2001; Gaut 2002) (Figure 9). The members of the *Poaceae* family have evolved from a common ancestor that lived 50 MYA to give rise to the major agricultural species for human and animal nutrition.
The separation of the *Panicoideae* sub-family (maize and sorghum) took place early in *Poaceae* postdating the separation between *Ehrhartoideae* (rice) and *Pooideae* that have diverged 46 MYA. Triticeae and *Panicoideae* sub-families were further separated to give rise to wheat and barley 13 MYA as well as maize and sorghum 9 MYA respectively (Figure 9).

3.3.2 Comparative genetic mapping

Despite large variation in genome size and organization in the grass family, genes have a tendency to be highly conserved at the DNA sequence level. Particularly, orthologous genes show a high conservation in the coding regions between distantly related species and an overall high gene conservation (exons and introns) between closely related species (Dubcovsky et al. 2001; Chantret et al. 2004). The identification of orthologous sequences has led to the first comparisons of genetic maps between different grass species.
Initial comparisons have revealed that homoeologous wheat chromosomes as well as chromosomes of closely related species such as barley and diploid wheat are remarkably conserved at the macro-level (Devos et al. 1993; Dubcovsky et al. 1996). Comparative studies among distantly related grass species such as between maize, rice and wheat species(Ahn et al. 1993; Kurata et al. 1994), between rice and Triticeae species (*T. aestivum*, *T. tauschii* and *Hordeum vulgare*) and/or maize and oat (VanDeynze et al. 1995a; VanDeynze et al. 1995b) have also established a significant conservation of marker and gene order along chromosomes (i.e. colinearity, also called macro-colinearity). Finally, a consensus map aligning the genomes of eight different grass species was drawn using as a reference rice, the smallest genome in grass species (Gale and Devos 1998). Globally, the chromosomal organization is conserved in 30 large blocks that are differentially rearranged in each grass genomes (Figure 10). Colinearity studies at the macro-level between rice and Triticeae species were validated at the mega-base level by the study of fine-scale DNA marker order and high-resolution mapping in the genomes of rice and Triticeae (Dunford et al. 1995; Distelfeld et al. 2004; Li et al. 2004). Such conservation of the colinearity among species that diverged 77 MYA and for which the genome size can vary 40

times, has established rice as a model species to study genome evolution in grasses. The fact that important agronomic traits such as dwarfing (Peng et al. 1999) were found conserved at orthologous location between distant grass species has reinforced rice as reference genome for positional cloning in species with large genomes (Keller and Feuillet 2000).

However, the conservation of the gene order and orientation at the mega-base level remains critical for efficient utilization of a model species for positional cloning, for development of markers and for the identification of regions that might contain candidate genes for the trait of interest.

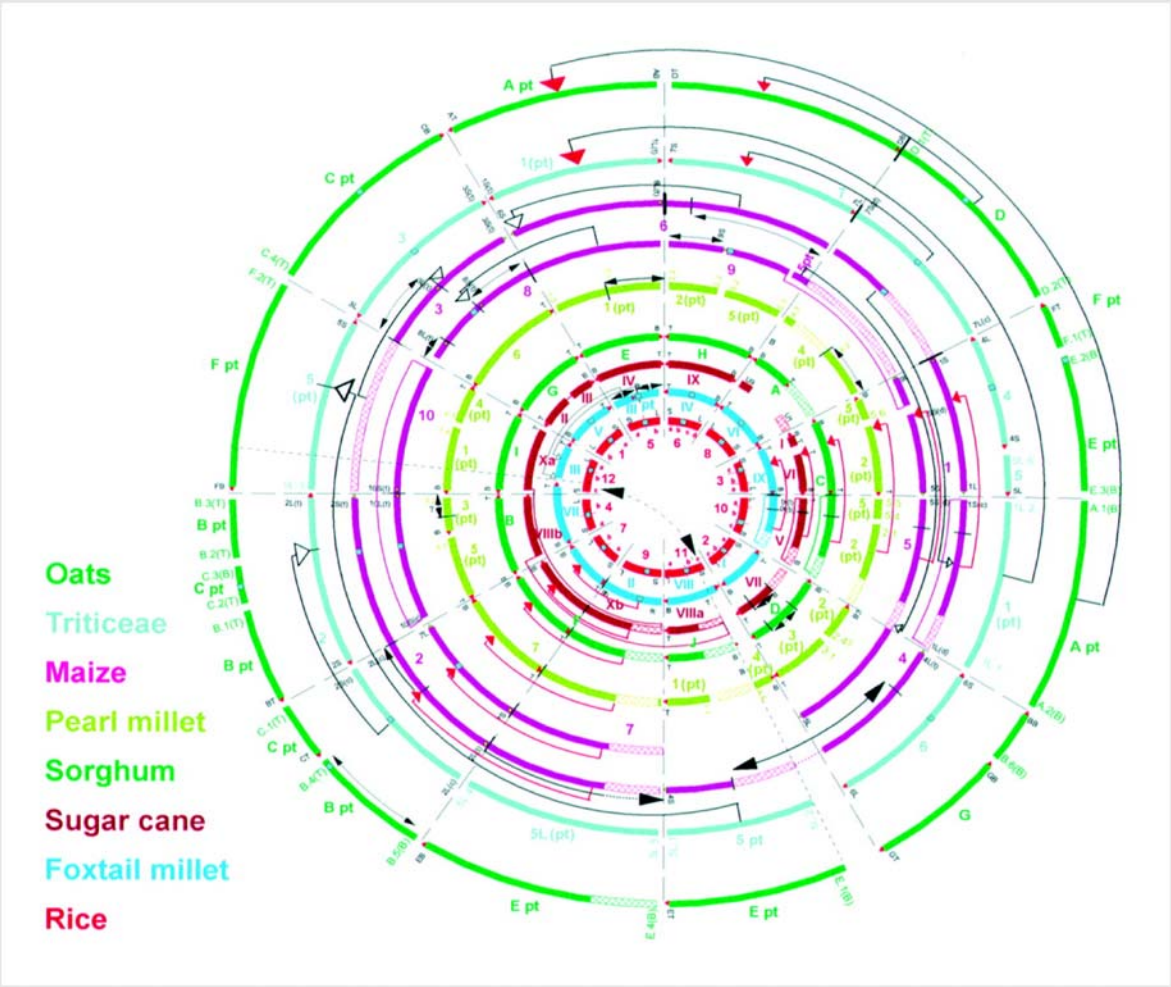

Figure 10 Consensus map of eight grass genomes
Each circle represents the chromosomal complement of a single grass genome. The circles are aligned, in the most parsimonious manner relative to rice and the arrows indicate the inversions and translocations, relative to rice. Locations of telomeres (black triangle) and centromeres (black square) are shown where known. Hatched areas indicate chromosome regions for which very little comparative data exist. L, long arm; S, short arm; T, top of chromosome; B, bottom of chromosome; and pt, part. From Gale and Devos, 1998.

The resolution of genetic mapping experiments remains low and cannot detect micro-rearrangements that might exist within apparently colinear regions. New strategies are then required to investigate the colinearity (also called micro-colinearity) at the sequence level, such as the whole genome comparative mapping and the direct comparisons of large stretches of genomic DNA

3.3.3 Sequence-based comparative genomics

The availability of the draft sequences of two domesticated rice cultivars *Oryza sativa* ssp. *japonica Nipponbare* and *Oryza sativa* ssp. *indica* 93-11 provided the basic data for large sequence comparison with other grass species (Goff et al. 2002). The recent physical assignment of 7,873 unique Triticeae ESTs to chromosome bins using deletion mapping experiments on hexaploid wheat deletion stocks provided a new source of data for comparison and validation of colinearity between rice and wheat genomes (Qi et al. 2003). The sequence-based comparative mapping between rice genomic sequences and the wheat deletion map shows an overall similarity to previous comparison at the genetic level (Figure 11).

However, a detailed comparison revealed a complex colinearity due to the high frequencies of rearrangements as well as numerous paralogous loci. This result suggested that the genome structures have significantly evolved since the radiation of grasses and may complicate the use of rice as a model for map based cloning (Sorrells et al. 2003a; La Rota and Sorrells 2004).

3.3.4 Micro-colinearity studies

Following the progress in isolation and sequencing of large stretches of genomic DNA in grass species, comparison of orthologous sequences has shown that even if the gross macro-colinearity is retained between different species, the conservation of the local structure can be incomplete, even between closely related species (Table 4).

Comparison between distantly related species in the grass family such as between rice and Triticeae or between rice and *Panicoideae* species has shown different mechanisms of micro-rearrangements (Figure 12). Some of them have little effect on the micro-colinearity and do not generate breaks: such events include local gene duplication and deletion of a tandemly duplicated gene generated and lost probably by unequal recombination mechanisms (Figure 12, A2, A3) (Dubcovsky et al. 2001; Ramakrishna et al. 2002; Song et al. 2002). The rapid divergences of the

size and composition of intergenic regions do not perturb the micro-colinearity as well. Such differences are due to a differential content of transposable elements, mainly LTR retrotransposons, in large genomes such as maize, wheat and barley (Figure 12, A1).

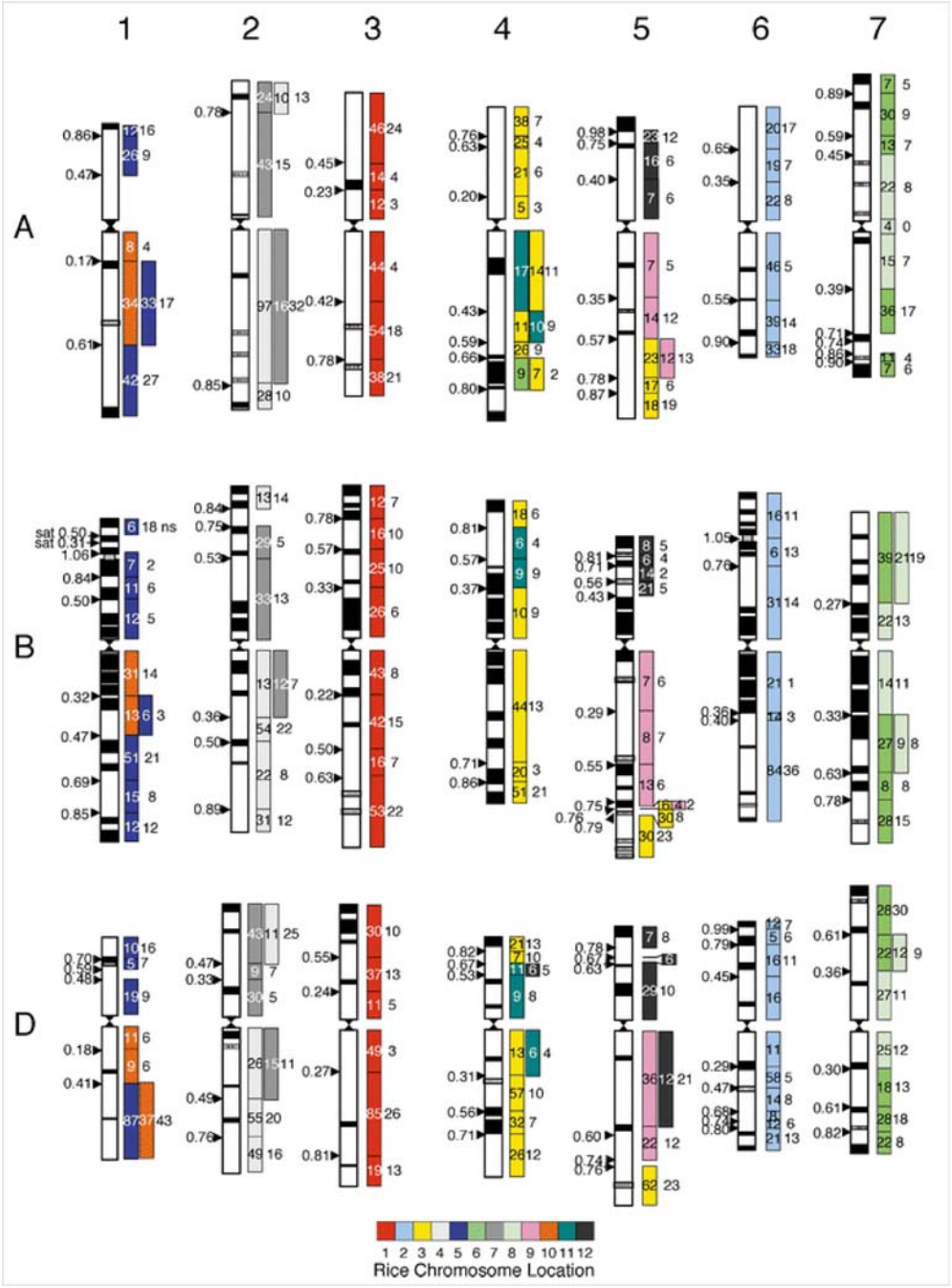

Figure 11 Wheat-rice genome relationships
The position of deletion breakpoints are indicated on the left of chromosomes as well as the FL (fraction length) values. Boxes on the right represent the deletion bins and are color-coded according to the most significant syntenic associations to rice chromosomes of wheat ESTs mapped to that bin. The number of those matches is indicated inside the colored rectangles while the number of matches to all other rice chromosomes is adjacent. Bins with non-significant associations or insufficient data are omitted. From La Rota and Sorrells (2004).

Mechanisms producing gene movements have a more important impact on the micro-colinearity and are responsible for numerous exceptions of the conservation of the colinearity. Gene inversions, deletions and translocations can create a complex mosaic conservation of the micro-colinearity between two orthologous regions (Bennetzen and Ramakrishna 2002; Feuillet and Keller 2002; Gaut 2002; Li and Gill 2002; Song et al. 2002). The rapidly evolving nature of certain genes such as disease resistance genes can also perturb the micro-colinearity in such loci (Keller and Feuillet 2000). Exceptions to colinearity were also reported even between closely related species.

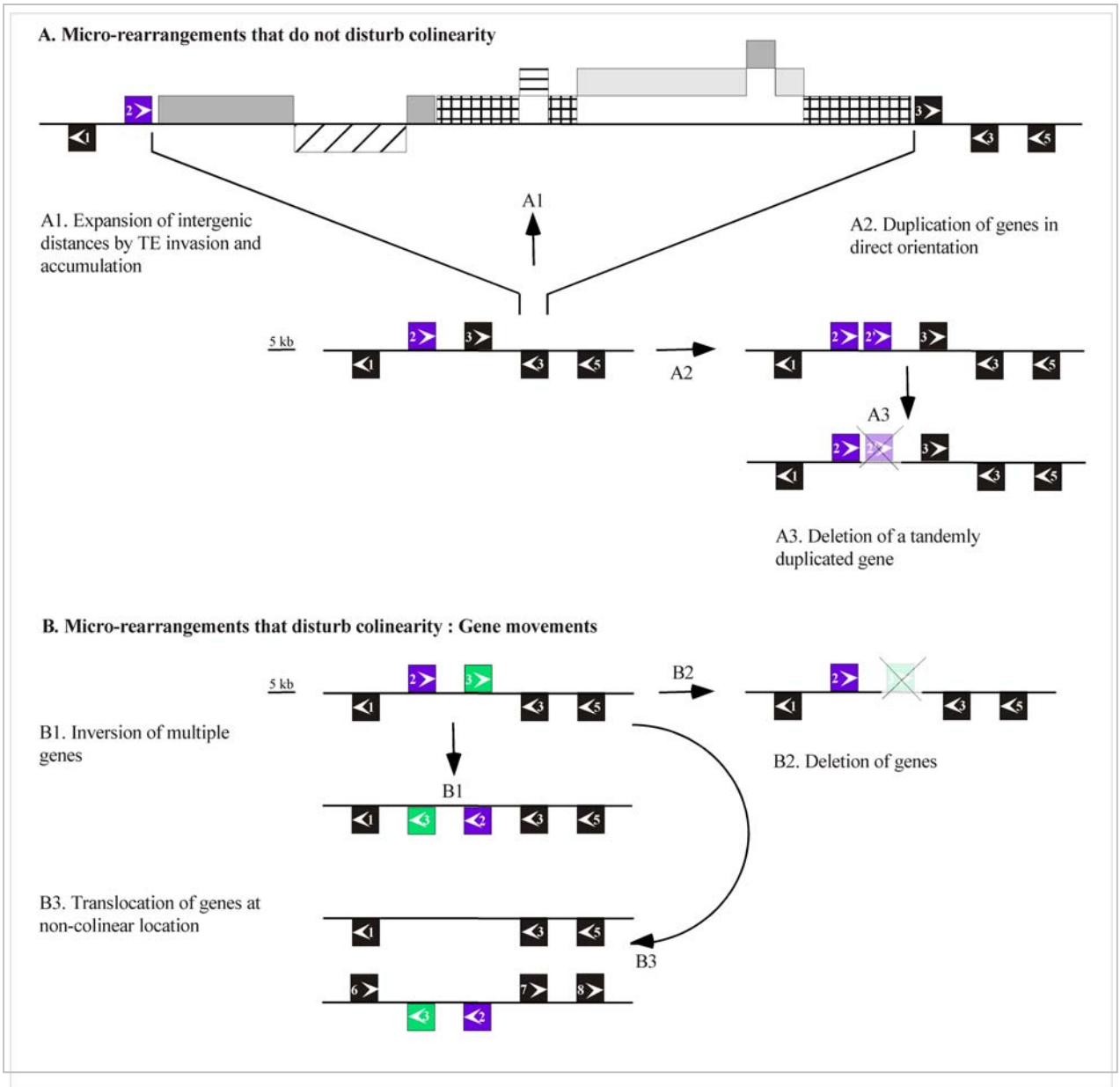

Figure 12 Micro-rearrangements observed at the sequence level between orthologous regions of grass genomes

Comparisons between different *Triticum* species at orthologous locations that diverged 1-3 MYA have shown a rapid divergence through the action of TEs (Wicker et al. 2003b). At the intra-species level, more dramatic differences have been reported in different maize genotypes (Fu and Dooner 2002; Song and Messing 2003).

There, in addition to a different TEs composition, gene movements have created breaks in the micro-colinearity. In contrast, comparisons between the two sequenced rice sub-species (*japonica* and *indica*) separated less than 1 MYA, have found a high conservation of the micro-colinearity with the exception of few TEs insertions and gene movements (Han and Xue 2003).

Altogether these studies indicate that the macro-colinearity does not necessarily reflect a good micro-colinearity, complicating the use of model species for map-based cloning but providing a good opportunity to study the evolution of genomes.

3.4 Aim of the study

The main objectives of this study were first to annotate molecular structures in Triticeae genomic sequences and to perform comparative genomic analyses between Triticeae and rice genomes in order to investigate the mechanisms of genome evolution in the Triticeae tribe. In the first part of the thesis, I will describe the discovery of an unusual local accumulation of Class II transposon CACTA types by a fine annotation of 427 kb of genomic sequences. The structure, organization and evolution of these elements as well as their impact on the wheat genome have been investigated. In the second part, I will describe an *in silico* comparative genomic analyses between 1.1 Mb of physical data at the distal part of the short arm of chromosome 1A in wheat and the rice genome. Comparison allowed to discover new mechanisms of the grass genome evolution such as ancestral large scale duplications in the rice genome. In the third part, the extent of such duplications in rice and their impact in the Triticeae genome evolution will be investigated. Finally, in the last part of the thesis, comparative genomic analyses between barley, wheat and rice allowed to discover additional evolutionary mechanisms that have shaped the Triticeae genomes.

Table 4 List of different micro-colinearity studies in grass species

	Locus	Species	Number of species	References	Colinearity	Rearrangements that do not disturb colinearity	Rearrangements that disturb colinearity
Interspecies comparisons	*adh1/adh2*	maize-rice-sorghum	3	(Tikhonov et al., 1999)	Conserved	Expansion of intergenic distances in maize due to transposable elements	Gene movements: deletion/insertion of genes
	adh1/adh2	maize-rice	2	(Tarchini et al., 2000)	Partially conserved	ND	Transposition of adh1 gene
	adh1/adh2	maize-maize-rice-sorghum	4	(Ilic et al., 2003)	Partially conserved	Tandem duplication in rice	Gene movements: gene insertion in the ancestor of maize and sorghum, insertion of two more genes in sorghum, gene loss in maize
	sh2/a1	maize-rice-sorghum	3	(Chen et al., 1997; Chen et al., 1998)	Conserved	Duplication of a1 in sorghum, expansion of intergenic distances in maize	ND
	sh2/a1	barley-maize-rice-sorghum-wheat	5	(Li and Gill, 2002)	Partially conserved	Duplication of X1, expansion of intergenic distances in wheat	Gene movements: translocation of two genes in Triticeae (sh2-X1)
	Ha	rice-diploid wheat	2	(Chantret et al., 2004)	Conserved	Duplication of gene 1a, expansion of intergenic distances in wheat due to TE	ND
	HMW-glutein	barley-wheat (D)	2	(Gu et al., 2004)	Conserved	Duplication of globulin and glutenin genes in wheat, variation of intergenic distances due to TE insertions	ND
	LMW-glutenin	diploid wheat-tetraploid wheat	2	(Wicker et al., 2003b)	Limited conservation, few genes are present due to high insertion of TE	Duplication of resistance genes, tandem duplication of large segments, TE variation	ND
	HMW-glutein	Tetraploid wheat (B)-diploid wheat (D)	2	(Kong et al., 2004)	Conserved	Variation of intergenic distances due to TE insertions	ND
	Lrk	barley-maize-rice-wheat	4	(Feuillet and Keller, 1999)	Partially conserved	Deletion/duplication of tandem genes	Gene movements: duplication/translocation of genes from chromosome 1 to 3
	Rph7	barley-rice	2	(Brunner et al., 2003)	Partially conserved	Inversion of gene HvHGA4, tandem duplication, expansion of intergenic distances in barley due to TE	Gene movements: Insertion of six additional genes in barley
	vrn1	barley-rice-sorghum-wheat	4	(Ramakrishna et al., 2002; Yan et al., 2003)	Conserved	Tandem duplication in the ancestor of Triticeae, expansion of distance due to TE	Inversion of gene 2 in barley, insertion and amplification in rice of 48 small nucleolar RNA genes
	Rp1	Maize and sorghum	2	(Ramakrishna et al., 2002)	Partially conserved	Numerous duplications of R genes, truncated copies of R genes	Gene movements: presence of non colinear genes in maize
	WG644	barley-rice	2	(Dubcovsky et al., 2001)	Conserved	Expansion of intergenic distances due to TE	Inversion of gene 2
	WG644	barley-rice-wheat	3	(SanMiguel et al., 2002)	Conserved	Expansion of intergenic distances due to TE	Inversion of gene 2
	chr4	rice (indica)-rice (japonica)	2	(Han and Xue, 2003)	Conserved	Limited insertion of TE however japonica > indica due to TE, a few gene tandem duplication	A few gene movements
	LR10	rice-diploid wheat (A)	2	(Guyot et al., 2004)	Partially conserved	Variation of intergenic distance due to TE insertion	Two resistance genes were not present in rice, inversion of two genes (ACT-CCF)
	Zein	maize-rice (indica)-rice (japonica)-sorghum	4	(Song et al., 2002)	Limited conservation, mosaic organization of collinar and non-colinear genes	Expansion of intergenic distance in maize due to TE insertion	Gene movements: Zein cluster (missing in rice) and other genes
Intraspecies comparisons	*bz*	maize-maize	2	(Fu and Dooner, 2002)	Partially conserved (haplotype differences)	Variation of intergenic distance due to TE insertion	Gene movement (4 gene added)
	Zein	maize-maize	2	(Song and Messing, 2003)	Partially conserved (haplotype differences)	Variation of intergenic distance due to TE insertion	Gene movement

Chapter 4

CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements

Thomas Wicker, Romain Guyot, Nabila Yahiaoui, and Beat Keller

(2003)

Plant Physiology 132:52-63.

4 CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements

4.1 Abstract

In comparison with retrotransposons, which comprise the majority of the Triticeae genomes, very few class 2 transposons have been described in these genomes. Based on the recent discovery of a local accumulation of CACTA elements at the *Glu-A3* loci in the two wheat species *Triticum monococcum* and *Triticum durum*, we performed a database search for additional such elements in Triticeae spp. A combination of BLAST search and dot-plot analysis of publicly available Triticeae sequences led to the identification of 41 CACTA elements. Only seven of them encode a protein similar to known transposases, whereas the other 34 are considered to be deletion derivatives. A detailed characterization of the identified elements allowed a further classification into seven subgroups. The major subgroup, designated the "*Caspar*" family, was shown by hybridization to be present in at least 3,000 copies in the *T. monococcum* genome. The close association of numerous CACTA elements with genes and the identification of several similar elements in sorghum (*Sorghum bicolor*) and rice (*Oryza sativa*) led to the conclusion that CACTA elements contribute significantly to genome size and to organization and evolution of grass genomes.

4.2 Introduction

All genomes contain repetitive elements and in some species, such elements comprise the majority of the nDNA. Repetitive elements can be divided into two main groups: class 1 and class 2 elements. Class 1 elements (also called retrotransposons) replicate via an mRNA intermediate that is reverse transcribed into DNA and integrated somewhere else in the genome. Retrotransposons contribute a large fraction to the total genomic DNA of plants with large genomes such as wheat, barley (*Hordeum vulgare*), or maize (*Zea mays)* (SanMiguel and Bennetzen 1998; Shirasu et al. 2000; Wicker et al. 2001; SanMiguel et al. 2002). Class 2 elements or transposons move via a DNA intermediate, which means that the elements are excised from the genome and integrated elsewhere. Excision and reintegration require an enzyme known as transposase. Transposons have been subdivided into several families. One of them, called the CACTA family, received its name because it is flanked by inverted repeats that terminate in a conserved CACTA motif. *En-1* (also known as Suppressor-mutator or *Spm*) from

maize was the first CACTA element that was analyzed at the molecular level (Pereira et al. 1986). *En/Spm* elements are present as autonomous elements that encode the proteins necessary for their transposition and deletion derivatives, which are nonautonomous. The nonautonomous elements depend for their transposition on enzymes encoded by the autonomous copies (Bennetzen 2000). Active CACTA elements were isolated and characterized from a variety of species including *CAC1* from Arabidopsis (Miura et al. 2001). *PsI* from petunia (*Petunia hybrida*) (Snowden and Napoli 1998), *Tdc1* from carrot (*Daucus carota*) (Ozeki et al. 1997), *Tam-1* from snapdragon (*Antirrhinum majus*) (Nacken et al. 1991), *Tpn1* from Japanese morning glory (*Ipomoea nil*) (Inagaki et al. 1994), and *Candystripe1* from sorghum (*Sorghum bicolor*) (Chopra et al. 1999). *Candystripe1* is believed to be a nonautonomous element because it does not encode a protein similar to known transposases.

The terminal regions of all identified CACTA elements show a similar sequence organization. They are flanked by short terminal inverted repeats (TIRs) of 10 to 28 bp in size that terminate in the CACTA motif. These serve as recognition sequences for the transposase protein (Lewin 1997). In most cases, sequence conservation between the different families is limited to this short motif, which makes it virtually impossible to identify new elements based on the TIR sequences of known elements. In addition, CACTA elements contain sub-terminal repeats (TRs) that consist of 10- to 20-bp units that are repeated in direct and inverted orientation. As for the TIRs, these units also show no significant sequence conservation between different families. Therefore, CACTA transposons are difficult to identify and usually are only found because of the presence of a transposase-like protein. Diploid Triticeae spp. such as barley or *Triticum monococcum* have genome sizes of more than 5,000 Mb and contain approximately 80% of repetitive DNA (Smith and Flavell 1975; Bennett and Leitch 1995). This high percentage of repetitive sequences has so far prevented them from becoming the focus of large-scale genomic sequencing projects. In recent years, however, a number of bacterial artificial chromosome (BAC) clones from Triticeae spp. were completely sequenced and to date, approximately 1.6 Mb of large contiguous stretches of genomic sequences are publicly available. Analysis of these sequences revealed that a large fraction of the repetitive DNA is comprised of retrotransposons (Shirasu et al. 2000; Wicker et al. 2001; Rostoks et al. 2002; SanMiguel et al. 2002; Wei et al. 2002), whereas class 2 elements were identified only in very few cases (Dubcovsky et al. 2001; Feuillet et al. 2001; Wei et al. 2002). So far, only one CACTA transposon from Triticeae has been described in detail (*TAT-1*)

(Feuillet et al. 2001). Therefore, it was assumed that this element class is present in a very limited copy number in the Triticeae genomes.

Recent analysis of the *Glu-A3* loci in diploid and tetraploid wheat revealed the presence of 12 different CACTA transposons (Wicker et al. 2003b). Interestingly, only four of these elements encode transposase proteins similar to those of previously described transposons. Eight of the 12 transposons were apparently deletion derivatives because they have no obvious coding capacity. Five of the deletion derivatives were designated as small nonautonomous CACTA (SNAC) transposons because their small size (700 bp-1.5 kb) clearly distinguished them from all other identified elements. The other three deletion derivatives range in size from 5 kb up to 11.3 kb.

The objective of our study was to characterize the previously described CACTA elements from wheat and to identify new Triticeae elements present in the public databases. Here, we report the identification and characterization of 41 novel CACTA transposons from Triticeae. Our results indicate that this transposon class is present at a high copy number in the wheat genome and that a large number are deletion derivatives. Elements similar to the ones in Triticeae were found in rice (*Oryza sativa*) and sorghum, indicating that also these genomes contain a wide variety of CACTA elements.

4.3 Material and Methods

Southern Hybridization of High-Density BAC Filters

Two copies of Filter C from the *Triticum monococcum* BAC library (Lijavetzky et al. 1999) were incubated over night at 65°C with radioactively labeled *Probe512* and *Probe179*, respectively. The filters were washed three times for 20 min at 65°C in 0.5× SSC and 0.1% (w/v) SDS and exposed to BIOMAX MS films (Eastman-Kodak, Rochester, NY) overnight.

Database Mining and Sequence Analysis

Public databases and the database for Triticeae repetitive elements (TREP, http://wheat.pw.usda.gov/ITMI/Repeats) were screened with the BLASTN and BLASTX algorithms (Altschul et al. 1997). For the identification of TR sequences, a 127-bp consensus sequence was used as a query for BLASTN search (consensus TR sequence:

CACTACTAGGGAAAAGGCCT-
ACTAATAGCGCACCGGATTGCTACTAATGGCGCCCAGGGGTGCGCC-
ACTAGCGCTACCACGCCAGTACTATATCTTACTAATGGCGCACCAGG-
GTGGTATAAACCC). Detailed sequence analysis was performed with the GCG Sequence Analysis Software Package version 10.1 (Devereux et al. 1984) and by dot-plot analysis (program DOTTER) (Sonnhammer and Durbin 1995). Sequence alignments were done with the GCG programs BESTFIT and PILEUP. The multiple alignment of the TR sequences was done with PILEUP (gap creation penalty = 2, gap extension penalty = 0). Phylogenetic analysis was performed with ClustalW (Thompson et al. 1994). Distances between pairs of TRs were calculated using the neighbor-joining method. Confidence values for the nodes were calculated using 1,000 bootstraps. For efficient processing of large sets of sequences, programs were written using the language PERL. Identified transposons were named as follows: The name of the transposon is separated by an underscore from the address of the BAC clone or the GenBank accession number of the sequence in which the element was discovered. Copy numbers of individual elements from the same source sequence are separated from the name by a hyphen.

4.4 Results

Identification of CACTA Transposons by BLAST Search and Dot-Plot Analysis

Because only a minority of the CACTA transposons was expected to actually encode a transposase-like protein, a first approach for the identification of new elements was based on their TR sequences. The TR regions that contain the TIRs and the sub-TRs usually have a size of 200 to 500 bp. In this study, the term "element with complete ends" was used for elements in which both TIRs contain an intact CACTA motif and are flanked by a 3-bp target site duplication. They were distinguished from elements truncated by deletions or elements with damaged ends (referred to as "truncated elements").

Ten of the 12 CACTA elements with complete ends identified on the *Glu-A3* contigs (Wicker et al. 2003b) showed conserved sequence motifs within their TR regions. These 10 elements, the previously described *TAT-1* (Feuillet et al. 2001) and another recently identified CACTA element from barley (*Caspar_AF521177-1*) (Brunner et al. 2003) were used to derive a 127-bp TR consensus sequence. This was used as a query sequence for a BLAST search of public databases

and the database for Triticeae repetitive sequences (Triticeae REPeat sequence database, http://wheat.pw.usda.gov/ITMI/Repeats) (Wicker et al. 2002). Ten new CACTA elements were found in genomic DNA sequences from Triticeae, six of which are elements with complete ends, whereas the other four were either truncated or only partially covered by the sequence deposited in the databases. In addition, nine Triticeae expressed sequence tags (ESTs) that contain TR sequences were identified. The presence of TR sequences in ESTs was interpreted as the result of transposon insertions close to genes. These were distinguished from EST sequences of the actual transcripts of the coding sequences of transposase-like proteins (see below). The transcript of transposon genes starts some 100 bp downstream of the TR region; therefore, it does not include the TR sequences. For one EST (accession no. BF618436), BLASTX search revealed that the CACTA element has presumably inserted in the 3'-untranslated region. In two cases, the element was inserted into the coding region of a Gag-Pol polyprotein (accession nos. BJ247168 and BJ253225).

It was clear that the consensus TR would not identify CACTA elements that contain divergent TRs. Therefore, a second approach for the identification of new elements was based on their structural similarity rather than on sequence conservation: The subterminal direct and inverted repeats displayed a specific pattern when the transposon sequence is plotted against itself with dot plot (Sonnhammer and Durbin 1995). The example in Figure 1 shows a dot plot of an SNAC element from *Triticum aestivum* (*Caspar_AF234649-1*). Typically, the short TIRs are immediately followed by a variable number of sub-TRs. In the case of the *Caspar_AF234649-1* element, they consist of direct and inverted repeats of a conserved 15-bp motif (CCTTTAGTCCCGGTT) that produce the characteristic "transposon signature." All transposons analyzed in this study contain sub-TRs within their TR sequences. Most elements contain two to six repeat units. Usually, the number of repeat units at one end differs from the number at the other end. A set of 15 publicly available large genomic Triticeae sequences and the two sequences from *T. monococcum* and *Triticum durum* (Wicker et al. 2003b) were collected in a local database with a total size of 1.9 Mb. This database was hereafter screened by dot plot for the occurrence of transposon signatures. This second approach led to the identification of six further CACTA elements with complete ends from genomic sequences.

In total, the database mining resulted in the identification of 16 new Triticeae CACTA transposons from genomic sequences and nine from EST sequences. None of the 16 new

elements found in genomic sequences had been annotated as such. It is likely that they were not recognized because none of them encodes a transposase protein. As it was previously described for retrotransposons in Triticeae, the CACTA elements were often found as nested insertions in other class 1 or class 2 elements.

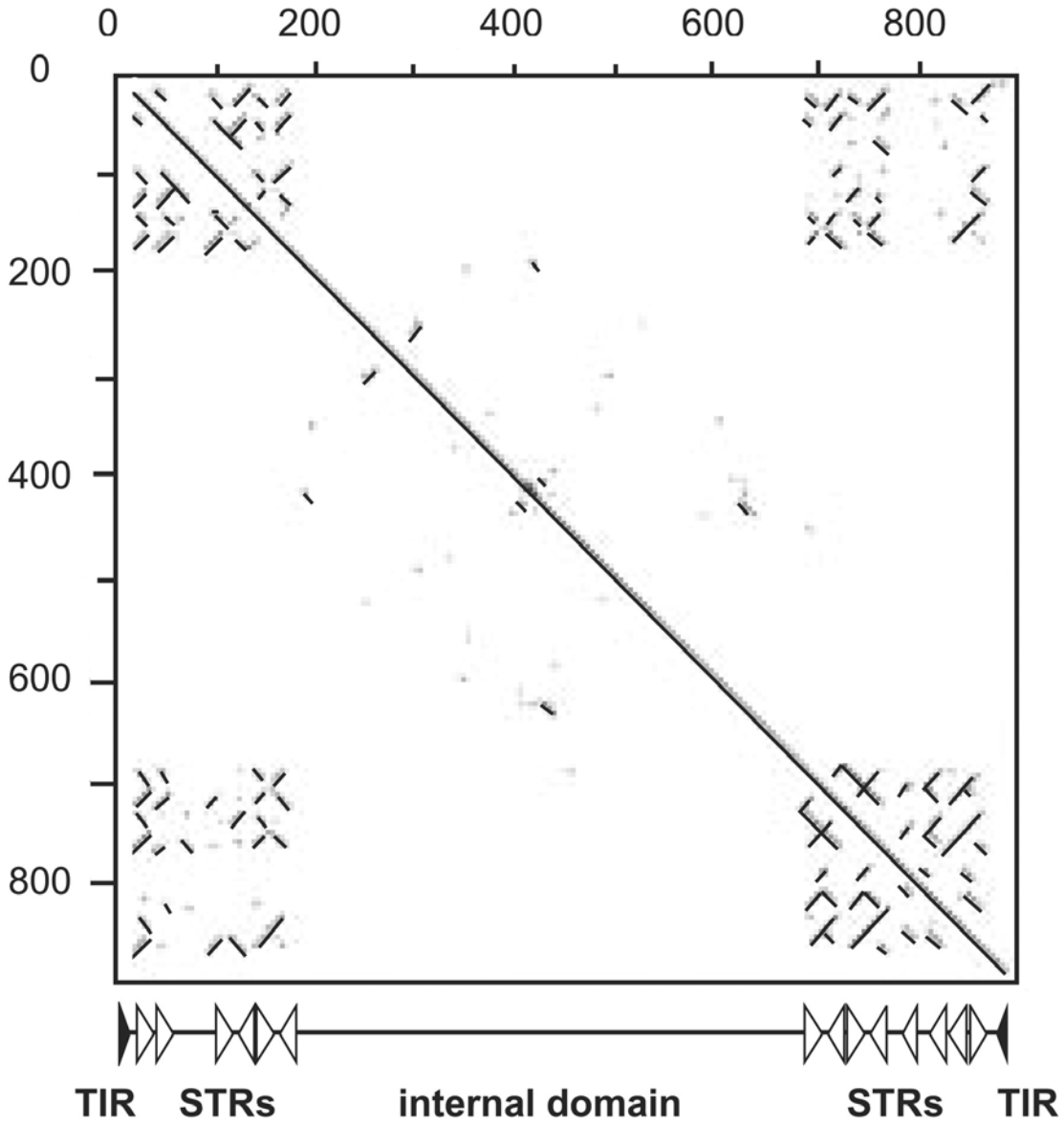

Figure 1 Dot plot of an SNAC transposon
The sequence of *Caspar_AF234649-1* is graphically compared with itself. The main diagonal line corresponds to the 100% match when the sequence is plotted against itself. Direct repeats are lines parallel to the diagonal line, and inverted repeats are displayed as lines perpendicular to the diagonal line. The sub-TRs produce a very specific pattern ("transposon signature") that can be easily recognized. The structure of the transposon is depicted below the dot plot. Black triangles, TIRs; white triangles, sub-TRs (STR).

Two additional elements (*Jorge_TREP766* and *Caspar_TREP788*) were kindly provided by Dr. Jorge Dubcovsky (University of California, Davis) and Dr. Nils Stein (Institute of Plant Genetics

and Crop Plant Research, Gatersleben, Germany), respectively. Together with the initial 12 elements, *TAT-1* (Feuillet et al. 2001) and *Caspar_AF521177-1* (Brunner et al. 2003), a total of 32 CACTA elements from genomic sequences are now available. Twenty-six of them are elements with complete ends in which both TIRs are present and a 3-bp target site duplication could be identified. All elements were collected in a local database and subsequently submitted to the TREP database (accession nos. TREP746-TREP788; http://wheat.pw.usda.gov/ITMI/Repeats). The names and origins of the identified elements are summarized in Table I.

The exact start and end positions of all identified elements in their source sequences are provided as supplemental material (see supplemental Table III at www.plantphysiol.org). As reference sequences, the previously described elements *En-1* from maize (Pereira et al. 1986), *Tam-1* from snapdragon (Nacken et al. 1991), *Candystripe-1* from sorghum (Chopra et al. 1999), and an additional CACTA element from *Lolium perenne* (accession no. AY089999), which was found by keyword search in the EMBL database, were also included.

CACTA Transposons Can Be Classified Based on Their TR Sequences

Because the majority of the identified transposons have no apparent coding capacity and vary greatly in size, we decided to base their classification on the TR sequences, the only feature that all of them have in common. The 14 truncated elements contain only one intact TR each, whereas from the 26 element with complete ends, both TRs could be used. The total 66 TR sequences from Triticeae transposons were used for a multiple sequence alignment. The alignment was done with the terminal 200 bp of the elements. A phylogenetic analysis of the multiple sequence alignments allowed the classification of the TR sequences into seven distinct clades (Fig. 2A). Sequence conservation between members of different families is restricted basically to the terminal 20 to 30 bp containing the CACTA motif. The major group containing 28 TR sequences was designated the "*Caspar* " family. One exclusive feature of the *Caspar* family is that the TR starts with a CACTAGT motif, whereas all others start with CACTAC(A/T). Three additional main families were designated *Balduin*, *Mandrake*, and *TAT-1*.

Table I List of all identified Triticeae CACTA transposons

The elements are sorted according to their classification into families. The family name is given first. The source sequence (e.g. BAC clone address, GenBank no., or TREP accession no.) follows the name after an underscore. For elements from genomic sequences, a number for the individual copy of an element is indicated after a hyphen. SNAC, Small nonautonomous CACTA. The reference refers to the researcher who published the sequence in which the element was found.

Name	Size (bp)	Organism	Comment	Reference
Caspar_107G22-1	1,305	T. durum	Complete ends, SNAC	Wicker et al. (2003)
Caspar_107G22-2	5,096	T. durum	Complete ends	Wicker et al. (2003)
Caspar_107G22-3	807	T. durum	Fragment	Wicker et al. (2003)
Caspar_18B1-1ab	9,913	T. monococcum	Complete ends	Wicker et al. (2003)
Caspar_426K20-2	1,371	T. monococcum	Complete ends, SNAC	Wicker et al. (2003)
Caspar_453N11-1ab	12,664	T. monococcum	Complete ends	Wicker et al. (2003)
Caspar_AF234649-2	750	T. aestivum	Complete ends, SNAC	O.D. Anderson, C. Hsia, and V. Torres (unpublished data)
Caspar_AF325197-1	1,483	T. aestivum	Fragment	Feuillet et al. (2001)
Caspar_AF325198-1	1,53	T. aestivum	Complete ends, SNAC	Feuillet et al. (2001)
Caspar_AF427791-1b	12,59	Barley	Complete ends	Wei et al. (2002)
Caspar_AF446141-1	6,255	Aegilops tauschii	Complete ends	Brooks et al. (2002)
Caspar_AF474072-1b	10,54	Barley	Complete ends	Rostoks et al. (2002)
Caspar_AF474373-1b	12,446	Barley	Complete ends	Rostoks et al. (2002)
Caspar_AF474373-2b	7,376	Barley	Complete ends	Rostoks et al. (2002)
Caspar_AF521177-1ab	9,105	Barley	Complete ends	Brunner el al. (2003)
Caspar_AL816166	284	T. aestivum	Fragment (EST)	I. Wilson, R. Beswick, S. Shepherd, G. Barker, J. Parker, P. Owen, D. Edwards, J. Coghill, M. Holdsworth, J. Lenton et al. (unpublished data)
Caspar_AL816390	341	T. aestivum	Fragment (EST)	I. Wilson, R. Beswick, S. Shepherd, G. Barker, J. Parker, P. Owen, D. Edwards, J. Coghill, M. Holdsworth, J. Lenton et al. (unpublished data)
Caspar_BF618436	143	Barley	Fragment (EST)	R. Wing, T.J. Close, A. Kleinhofs, R. Wise, D. Begum, D. Frisch, Y. Yu, D. Henry, M. Palmer, T. Rambo et al. (unpublished data)
Caspar_TREP770b	8,334	A. tauschii	Complete ends	P. Langridge (unpublished data)
Caspar_TREP771	3,296	A. tauschii	Fragment	P. Langridge (unpublished data)
Caspar_TREP788ab	11,627	Barley	Complete ends	N. Stein (unpublished data)
Caspar_X63357	569	Barley	Fragment	Fernandez et al. (1993)
Caspar_Z66528-1	331	Barley	Fragment	A. Molina, I. Diaz, F. Garcia-Olmedo (unpublished data)
Mandrake_107G22-1	930	T. durum	Complete ends, SNAC	Wicker et al. (2003)
Mandrake_426K20-1	898	T. monococcum	Complete ends, SNAC	Wicker et al. (2003)
Mandrake_AF234649-1	1,119	T. aestivum	Complete ends, SNAC	O.D. Anderson, C. Hsia, and V. Torres (unpublished data)
Mandrake_AF446141-1	3,411	A. tauschii	Complete ends	Brooks et al. (2002)
Mandrake_BG309793	736	Barley	Fragment (EST)	R. Wing, T.J. Close, A. Kleinhofs, R. Wise, D. Begum, D. Frisch, Y. Yu, D. Henry, M. Palmer, T. Rambo et al. (unpublished data)
TAT-1_AF325196-1a	7,19	T. aestivum	Complete ends	Feuillet et al. (2001)
TAT-1_107G22-1	3,158	T. durum	Fragment	Wicker et al. (2003)
TAT-1_AF427791-1	9,765	Barley	Complete ends	Wei et al. (2002)
TAT-1_BJ247168	414	T. aestivum	Fragment (EST)	Y. Ogihara, K. Murai (unpublished data)
TAT-1_BJ253225	514	T. aestivum	Fragment (EST)	Y. Ogihara, K. Muirai (unpublished data)
Balduin_453N11-1a	9,841	T. monococcum	Complete ends	T. Wicker et al. (unpublished data)
Balduin_BJ218039	554	T. aestivum	Fragment (EST)	Y. Ogihara, K. Murai (unpublished data)
Balduin_BQ469151	376	Barley	Fragment (EST)	H. Zhang, W. Weschke, W. Michalek, N. Stein, A. Graner (unpublished data)
Balduin_BQ661452	376	Barley	Fragment (EST)	W. Michalek, W. Weschke, K.-P. Pleissner, A. Graner (unpublished data)
Jorge_AF326781-1	16,497	T. monococcum	Complete ends	Wicker et al. (2001)
Jorge_TREP766	14,733	T. monococcum	Complete ends	J. Dubcovsky (unpublished data)
Enac_453N11-1	11,361	T. monococcum	Complete ends	T. Wicker et al. (unpublished data)
Isaac_107G22-1a	13,135	T. durum	Complete ends	T. Wicker et al. (unpublished data)

a Encodes a transposase. b Encodes a CTG-2 protein.

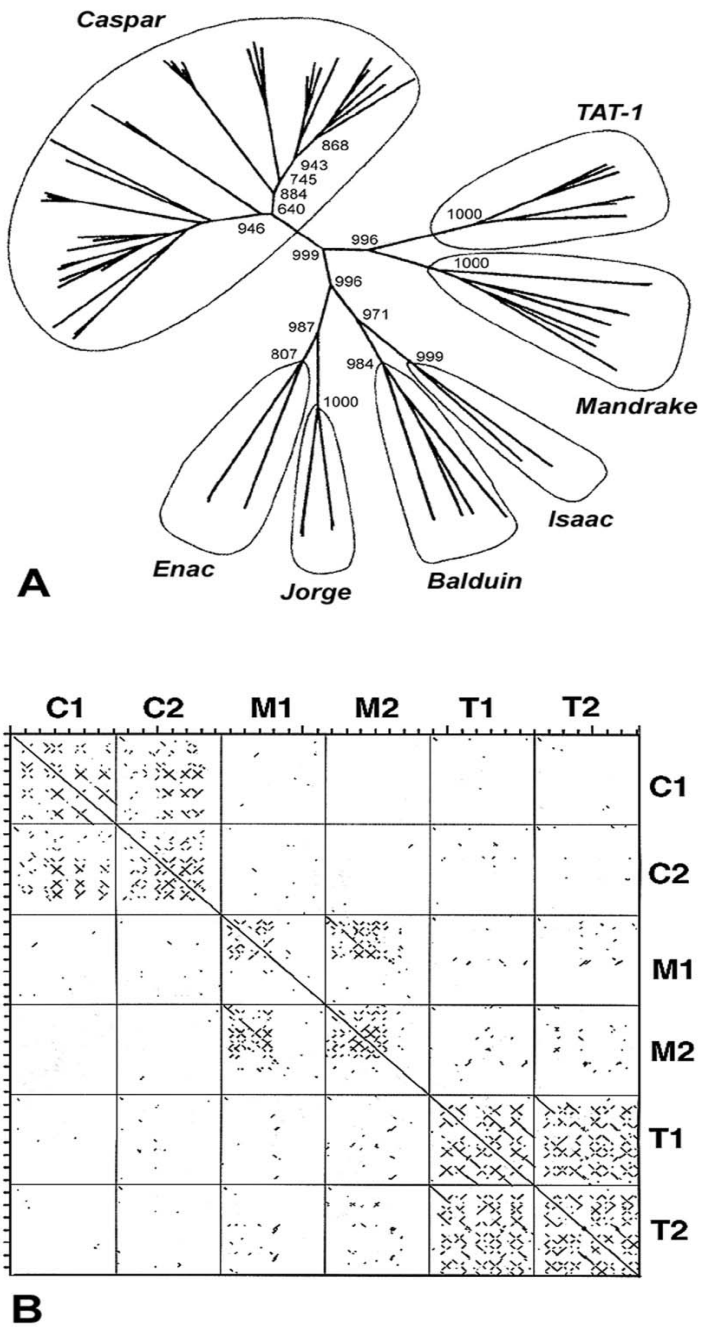

Figure 2 Classification of Triticeae CACTA transposons based on their TR sequences. A, Classification based on multiple sequence alignment. The bootstrap values for the seven main families and the major subfamilies are indicated at the nodes of the tree. The TR sequences were aligned with PileUp and analyzed with the neighbor-joining method. B, Classification based on the dot-plot pattern. An array of six TR sequences is plotted against itself. The terminal 300 bp from one TR of an element are used for the array. Only TRs from elements belonging to the same family display the characteristic transposon signature when compared with each other, whereas TRs from different families display no signature when they are plotted against one another. C1 and C2, TRs of *Caspar* elements. M1 and M2, TRs of *Mandrake* elements. T1 and T2, TRs of *TAT-1* elements.

Further similarities were discovered between *Jorge_TREP766* and the previously described unclassified *XB* element (Wicker et al. 2001), which was called thereafter *Jorge_AF326781-1*.

The TR sequences of *Enac_453N11-1* and *Isaac_107G22-1* are unique because they show no similarity to any of the other elements and groups in separate clades (Fig. 2A).

To test this classification, a second approach for classification was based on the similarity of TR sequences displayed by dot-plot analysis: TRs from members of the same family display the characteristic transposon signature, whereas TRs of elements from different families show no signature. The terminal 300 bp of one TR from each element was used to generate a large array, which was then compared against itself by dot plot. Examples for dot-plot alignments of three different families are displayed in Figure 2B. In this approach, the classification into seven groups as it was obtained by the multiple sequence alignment could be confirmed for all elements. The results of the two classification approaches are summarized in Table I.

The CACTA Family Comprises Full-Length Elements and a Wide Variety of Deletion Derivatives

To investigate the range of diversity in size and sequence organization among members of the CACTA family, only the 26 elements with complete ends were used.Truncated elements were excluded because it is not possible to determine their actual size and coding capacity. Seven of the 26 elements with complete ends encode a transposase protein (Table I). However, all seven do not encode functional proteins because they all contain frameshifts or in-frame stop codons within their coding region (see below). In this study, we refer to elements that encode a transposase protein as "full-length elements," even if the coding region of the transposase protein is apparently defective. Four of the seven elements encode a second protein (which we refer to as *CTG-2*) in addition to the transposase. The *CTG-2* coding gene was only found in the members of the *Caspar* family (see below). All identified full-length elements are large in size, ranging from 9.9 up to 13.1 kb.

The other 19 CACTA transposons are considered to be deletion derivatives that have lost some or all of their coding capacity and depend for their transposition on enzymes encoded elsewhere in the genome. These deletion derivatives vary drastically in size: At one end of the spectrum, there are seven SNAC transposons that encode no proteins and range in size from 750 bp to 1.5 kb. The TR regions of these seven SNAC elements have sizes of 200 to 300 bp and are separated by an internal domain.

Three SNAC elements belonging to the *Caspar* family (*Caspar_107G22-1*, *Caspar_426K20-2*,

and *Caspar_AF325198-1*) plus a fragment of a putative SNAC element (*Caspar_107G22-3*) contain a 64-bp region that is 75% to 81% identical to a part of the 5S rDNA gene (120 bp) from *T. monococcum* (accession no. Z11461). This region is embedded in an approximately 400-bp region that is more strongly conserved than the rest of the elements. In the 400-bp region, all four are 91% to 95% identical, whereas their overall sequence identity is 79% to 91%. The 5S derivative conserved in the four elements corresponds to the internal RNA polymerase III promoter that is involved in the recruitment of transcription factors. It includes the highly conserved motifs Box A, IE, and Box C (Cloix et al. 2000). In addition, three of the four elements contain a 191-bp region that is 63% to 90% identical to the spacer region of the 5S rDNA gene in *Hordeum cordobense* (accession no. AY034735). In total, 61 5S rDNA from barley gave strong BLASTN hits with this 191-bp region. The other three SNAC elements belong to the *Mandrake* family and show no obvious structure within their internal domain.

The 12 large deletion derivatives range in size from 3,411 bp up to 16.5 kb. Seven are members of the *Caspar* family, five of which encode a *CTG-2* protein. All seven large *Caspar* deletion derivatives contain regions of tandem repeated DNA (see below). The other five deletion derivatives do not contain any sequences similar to known repetitive elements or genes. They also do not contain obvious structures like direct repeats, which would explain their large size. The largest deletion derivative identified is *Jorge_AF326781-1*, which has a size of 16,497 bp.

Elements of the Caspar Family Encode a Transposase and a Protein of Unknown Function

Four *Caspar* elements (*Caspar_453N11-1*, *Caspar_18B1-1*, *Caspar_AF521177-1*, and *Caspar_TREP788*) gave strong BLASTX hits with numerous transposase-like proteins from rice and sorghum. The coding region for the transposase is located in the 5' region of the elements. All four are likely to be nonfunctional because they all contain frameshifts or in-frame stop codons within their coding regions. However, because they show a high degree of sequence conservation within the coding region of the transposase, a multiple sequence alignment allowed to determine at which positions frameshifts have to be introduced in an individual element to obtain a contiguous open reading frame. All four elements contain between one and three frameshifts and *Caspar_453N11-1* and *Caspar_TREP788* contain one and two in-frame stop codons, respectively. Comparison with transposase proteins from public databases helped to determine the positions of the putative start and stop codons. The four deduced transposase proteins have sizes

ranging from 1,044 to 1,122 amino acids and are 73% to 79% similar to one another. The coding region does not contain any introns. The four putative proteins are 68% to 74% similar to TNP2-like proteins from rice (accession no. Q9AUX7) and from sorghum (accession no. Q9XEQ1) but only 40% to 45% similar to the transposase of *En/Spm* (accession no. AAA66266). The transposase genes of *Caspar* elements are expressed as more than 30 ESTs from Triticeae corresponding to the transposase region were found in public databases.

Nine *Caspar* elements contain a coding region for a second protein we refer to as *CTG-2* (*Caspar* transposon gene 2). BLAST search of the *CTG-2* region revealed similarity to 12 hypothetical proteins from rice and one from sorghum. In contrast to the transposase, which is well conserved among the different *Caspar* elements, the *CTG-2* protein is highly variable. Therefore, it was difficult to predict a protein sequence. Based on sequence conservation between different *Caspar* elements and on the similarity to the proteins identified by BLASTX, putative protein sequences of eight *Caspar CTG-2* proteins were deduced. The proteins have sizes of 968 to 1,292 amino acids. In all cases, they consist of one large putative first exon, which varies strongly in size between the different copies. The differences are caused by a region that contains multiple repeats of short 3- to 30-bp units, and the number of repeat units differs in the different elements. This putative first exon is followed by five short exons (25-50 amino acids) that show a higher degree of sequence conservation. The exon/intron structure of the last five exons was determined by comparison with the amino acid sequences of the 12 hypothetical proteins from rice that were identified by BLASTX. The predicted exon/intron structure of *CTG-2* is strongly conserved in all analyzed elements. Eight ESTs similar to the *CTG-2* region were found in public databases, indicating that the *CTG-2* proteins are also expressed.

The predicted *CTG-2* protein sequences show no clear homology to previously described transposon proteins. A weak similarity to previously described proteins could be shown if sequences were aligned with the GCG program BESTFIT (Genetics Computer Group, Madison, WI), and gap creation and gap extension penalties were decreased to 4 and 1, respectively. Using these parameters, all *CTG-2* proteins are between 42% and 50% similar over most of their length to the TNP1 protein of *Tam-1* (accession no. CAA40554) and TNPA of *En/Spm* (accession nos. AAG17044). However, the sequence alignments contain a large number of gaps; therefore, one can only speculate that the *CTG-2* protein may represent a highly diverged homolog to TNP1 and TNPA.

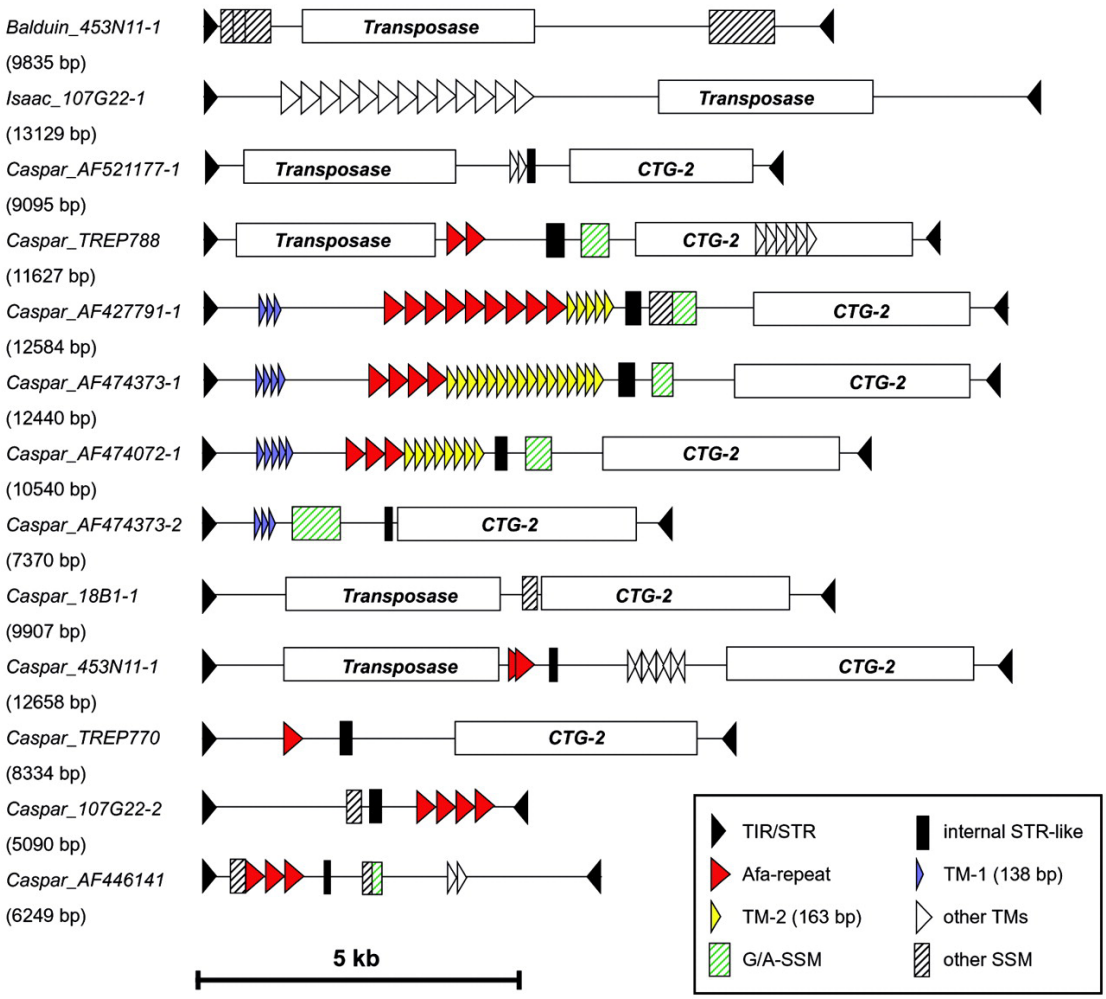

Figure 3 Repeat structures within different CACTA elements. Direct repeats larger than 100 bp are displayed as triangles. Repeat regions with shorter units are indicated as shaded boxes. TM, Tandem repeat; SSM, tandem repeats of short sequence motifs; STR, sub-TR.

CACTA Elements Contain Large Amounts of Low-Complexity DNA

Dot-plot analysis of the identified transposons revealed that several elements contain patterns of tandem repeats of variable length and sequence.

The repeated sequence units range in size from 2 to 30 up to 380 bp. A selection of 13 CACTA elements with complete ends that contain multiple different repeat structures were chosen for further analysis (Fig. 3). Eleven of them are members of the *Caspar* family, and the two others are *Balduin_453N11-1* and *Isaac_107G22-1*. SNAC transposons, the large deletion derivatives *Jorge_TREP766*, *Jorge_AF326781-1* and *Enac_453N11-1*, and truncated elements were excluded because they do not contain comparable repeat patterns. The repeat regions in *Balduin_453N11* and *Isaac_107G22* showed no similarity to each other or to the ones from the *Caspar* family, whereas nine of the 11 *Caspar* elements share common repeat units. A surprising finding was that

eight *Caspar* elements contain the previously described *Afa* repeats (Rayburn and Gill 1986; Nagaki et al. 1998). *Afa* repeats are a class of tandem repeats of approximately 340 bp in size that are believed to be present in all Triticeae spp. Their copy number, however, was shown to vary up to 100-fold in different Triticeae spp., and they were found in various, genome-specific locations in Triticeae genomes (Nagaki et al. 1998). Copy numbers of the *Afa* repeats in the identified *Caspar* elements range from one (*Caspar_TREP770*) to nine (*Caspar_AF427791*; Fig. 3). Two further repeat types (*TM-1* and *TM-2*) occur in three and four elements, respectively. In addition, most of the *Caspar* elements contain large regions (200-500 bp) of tandem repeated short sequence motifs (most often G/A-rich regions) and a region of 100 to 250 bp that is 70% to 85% identical to their sub-TRs (Fig. 3).

The tandem repeats within CACTA elements obviously can undergo rapid changes in copy number: Four *Caspar* elements from barley (*Caspar_AF427791-1*, *Caspar_AF474373-1*, *Caspar_AF474373-2*, and *Caspar_AF474072-1*) appear to be very closely related because they are approximately 92% to 95% identical on the DNA level. However, the most striking difference between them is the number of direct repeats (Fig. 3). *Caspar_AF427791-1*, for example, contains three copies of *TM-1*, nine Afa repeats, and five copies of *TM-2*, whereas *Caspar_AF474373-1* contains four *TM-1* units, four *Afa* units and 16 *TM-2* units. In contrast, *Caspar_AF747373-2* contains only four *TM-1* repeats but neither *Afa* nor *TM-2* repeats (Fig. 3).

The Caspar Family Is Present at a High-Copy Number in the Wheat Genome

The fact that the transposons of the *Caspar* family were found in several copies in the publicly available sequences suggested that these elements may occur very frequently in Triticeae genomes. To estimate the copy number of the *Caspar* transposons, one high-density filter (Filter C) from the *T. monococcum* BAC library (Lijavetzky et al. 1999) was hybridized with two different probes. One high density filter contains 18,432 BAC clones that cover approximately 0.4 genome equivalents. The first probe (*Probe512*) was chosen in the 5' region of the transposase-coding region of the *Caspar_453N11-1* element, and the second one (*Probe917*) covers the 3' region of *CTG-2* of *Caspar_453N11*. These two probes allowed the determination of how many elements contain both proteins and how many contain only one of the two. The hybridization pattern of both probes from a small region of filter C is shown in Figure 4. *Probe512* and *Probe917* identified 672 and 795 BAC clones, respectively, and 292 BACs gave

signals with both probes. These numbers were extrapolated to one genome equivalent (multiplied by 2.5). From these data, we estimate that the wheat genome contains a minimum of 2,900 copies of the *Caspar* elements. About 25% of them contain both the transposase and the *CTG-2* region. Approximately 950 copies contain only a transposase, and 1,250 copies contain only *CTG-2*. If one takes the average size of the nine *Caspar* transposons that encode one of the two proteins (10.5 kb), the roughly 3,000 *Caspar* elements might contribute approximately 0.6% to the *T. monococcum* genome. As shown above, many *Caspar* elements contain neither of the two proteins and are excluded from this estimate. It also has to be considered that the estimated copy number from the hybridization data was based on the assumption that each BAC clone that gave a signal contains only one *Caspar* element. Therefore, the actual number of *Caspar*-like transposons in the wheat genome might be considerably higher.

Caspar-Like Elements Are Also Frequently Found in Other Grass Genomes

The apparently high copy number of *Caspar* elements in Triticeae genomes inspired the search for similar elements in other grass genomes. Three BACs from rice and one from sorghum encoding the proteins that gave the strongest BLASTX hits with *CTG-2* from *Caspar* were screened for the presence of transposon-like sequences. In all four cases, an annotated transposase protein was found upstream of the protein that gave the BLASTX hit with *CTG-2*, but transposase and *CTG-2* were not annotated as belonging to the same element. In all four cases, *CTG-2* was annotated as a putative gene. The predicted exon/intron structure as it was annotated in the publicly available sequences differed slightly from our prediction of the structure of *CTG-2*. However, comparison with our predicted proteins from the Triticeae elements showed that that the same exon/intron structure also can be found in the elements from rice and sorghum, although the proteins from the different species were only about 46% to 50% similar to one another. Two proteins from rice BACs AP002484 and AP003020 and one from sorghum BAC AF114171 were deduced by applying our predicted exon/intron structure and used as query sequences for a TBLASTN search. The number of hits was striking: CTG-2_AP002484 and CTG-2_AP003020 gave 218 and 214 hits in rice, respectively, with E values below 3E-4. CTG-2_AF114171 identified five putative *CTG-2* proteins in sorghum (E value = 0.0).

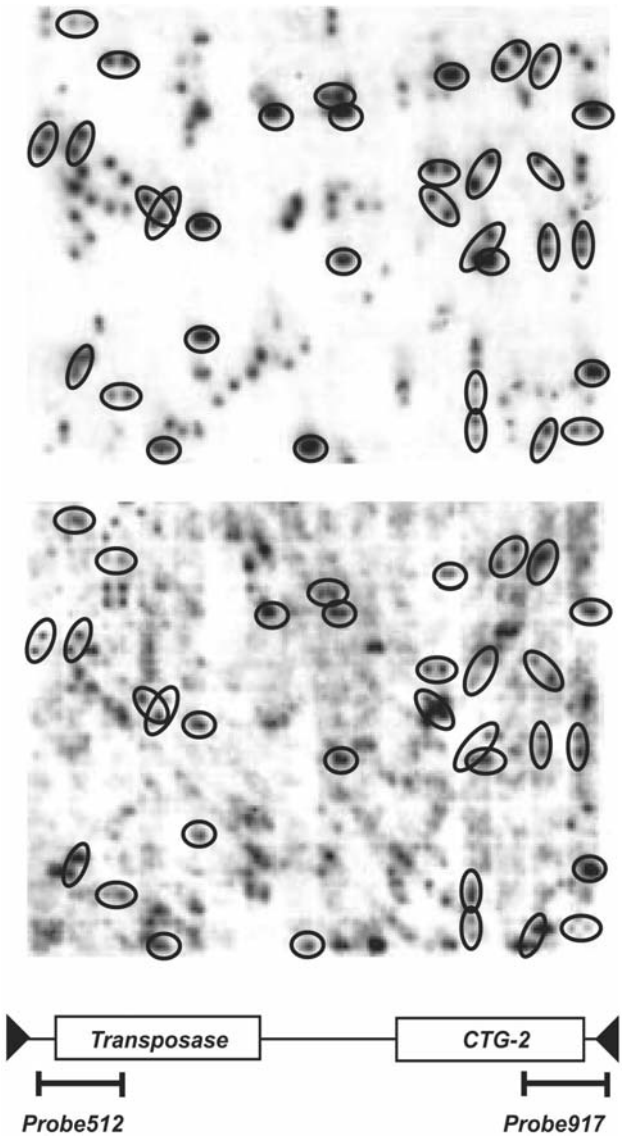

Figure 4 Estimation of the copy number of *Caspar* elements in the *T. monococcum* genome. One BAC filter was hybridized with two different probes corresponding to the transposase (top) and *CTG-2* regions (bottom) from *Caspar_453N11-1*, respectively. The fraction of the filter shown corresponds to approximately 3.3% of a genome equivalent. BAC clones that hybridized with both probes are indicated with circles.

Using dot plot, the actual borders of the elements on the rice and sorghum BACs were identified, and four *Caspar*-like elements with complete ends could be characterized. In addition, the four BAC clones were searched for further transposon signatures by dot plot, which led to the identification of two additional SNAC transposons (one from rice BAC AP002484 and one from sorghum BAC AF114171), both of which were not annotated. The positions of the elements on their respective BAC clones are shown in Table II.

All sequences identified in this way were used for a next round of BLASTN search against the

National Center for Biotechnology Information nonredundant database to obtain a rough estimate of the abundance of these elements in the rice and sorghum genomes. This search revealed the presence of a very high number of similar elements in the genomes of rice and sorghum, ranging from 493 hits for SNAC_AP002484-1 up to 824 hits for the CACTA element from rice BAC AP003020 that contains both a transposase and *CTG-2*. E values for all these BLASTN hits were below 3E-4. The CACTA element from sorghum BAC AF114171 identified four elements in sorghum (E value = 0.0). Because the focus of this study was not a complete survey of rice CACTA elements but to study their structure and sequence organization, we focused our attention on the isolation of a small number of elements with complete ends. The result of the database mining was a set of 18 CACTA elements from rice and six elements from sorghum. The precise location of all identified rice and sorghum elements on their source sequences is provided as supplemental material (Table III). Interestingly, only one additional element that encodes proteins was identified, and all others were SNAC transposons. None of the SNAC transposons had been annotated as such. These data suggest that the rice genome might contain a very large number of yet undiscovered CACTA elements and that the majority of them might be small nonautonomous elements. A very interesting finding in this context is *SNAC_AP003446-1* from rice, which at 274 bp is the smallest element identified in this study (Table II). It is the only element that does not contain an internal domain but consists exclusively of terminal and sub-TR sequences.

Table II Examples of CACTA elements from rice and sorghum Positions of the elements on the BAC clone are indicated

BAC Clone	Size (bp)	Positions	Organism
AC079029[ab]	10,544	51,199-64,657	Rice
AP002484[a]	10,545	15,795-26,339	Rice
AP002484[ac]	1,21	9,911-11,120	Rice
AP003020[a]	11,28	111,118-122,397	Rice
AP003446[c]	274	57,711-57,984	Rice
AF114171[a]	12,722	109,076-121,797	Sorghum
AF114171[ac]	1,716	82,698-84,413	Sorghum

[a] Used for the initial BLASTN search. [b] Contains insertions of repetitive elements. [c] SNAC transposon.

4.5 Discussion

Why Were the CACTA Elements in Triticeae Not Discovered Earlier?

The high density of CACTA elements observed at the *Glu-A3* loci from *T. monococcum* and *T. durum* was a fortunate constellation (Wicker et al., 2003). It allowed the characterization of a large number of elements belonging to different families and conclusions to be drawn about their general features and structures. The main reason why CACTA elements have remained undiscovered for so long is that not enough sequence data was available for the identification of these elements. From the handful of CACTA elements that were described so far in other species, only limited conclusions could be drawn as to what types of elements could be expected to be present in Triticeae. As we show in this study, sequence conservation at the DNA level is very low even between Triticeae elements and limited to the very TR regions among different grass species. A second reason for them being hidden so well is the unexpected finding that most CACTA elements are deletion derivatives and do not encode transposase proteins. Several elements containing the *CTG-2* proteins actually had been described before but due to the misleading BLASTX results had been interpreted as putative genes. In one case, the sub-TR structures flanking the *CTG-2* were interpreted as arrays of very small miniature inverted-repeat transposable elements (MITEs) (Wei et al. 2002).

CACTA Sequences in Grass Genomes Are Mainly Deletion Derivatives

All identified CACTA elements appear to be defective or nonautonomous because they either lack sufficient coding capacity, or their coding sequences are interrupted by frameshifts or in-frame stop codons. For *En/Spm* and *Ac/Ds* elements from maize, it was shown that numerous deletion derivatives exist that are only able to transpose in the presence of a functional element (Gierl and Saedler 1989). One can speculate that the initial autonomous *Caspar* transposon had a size of approximately 10 kb and encoded both a transposase and a *CTG-2* protein. During evolution of these elements, a large number of deletion derivatives were established, which themselves evolved and diverged further. Obviously, a large number of elements have lost their transposase region but have maintained the *CTG-2*, whereas other elements have lost both proteins and were reduced basically to their TR regions, which are in most cases separated by a

small internal domain (SNAC transposons). A possible final product of this tendency of size reduction is the *SNAC_AP003446-1* transposon from rice that does not even contain an internal domain but consists exclusively of TR sequences. Therefore, *SNAC_AP003446-1* might represent the "minimal transposon" that is reduced to its very basic functional components. All SNAC elements identified in this study contain both TIR and sub-TR sequences. This differentiates them from the previously described mobile element-like sequences, which also contain a conserved CACTA motif but only have TIR sequences (Hoshino et al. 2001). Both SNAC elements and mobile element-like sequences resemble MITEs (Bureau and Wessler 1994a), which are also considered to be nonautonomous elements.

However, during the evolution of nonautonomous elements, there was obviously no selection pressure that would favor smaller sized elements, as is illustrated by the numerous large elements such as *Jorge_AF326781-1*. An even more impressive example is the 23-kb *Candystripe1* transposon from sorghum. This CACTA element was shown to be active in sorghum, although it is also considered to be nonautonomous (Chopra et al. 1999). This concept can be expanded to other classes of repetitive elements. For example, the *Sabrina* retrotransposon (Shirasu et al. 2000) is one of the most abundant retroelements in Triticeae, but only few copies that actually encode a protein similar to reverse transcriptase were identified so far (SanMiguel et al. 2002; Wei et al. 2002). Thus, we conclude that nonautonomous repetitive elements are widely present in grass genomes and possibly include the majority of all mobile DNA sequences. Therefore, the Triticeae genomes may contain an enormous number of such nonautonomous elements, and many of them have not yet been discovered because they lack obvious coding sequences.

The Presence of Afa Repeats in Caspar Elements Explains Some of the Features of These Repeats But Also Raises New Questions

Because *Afa* repeats were found in several members of the *Caspar* family but never isolated outside of *Caspar* elements, we conclude that all *Afa* repeats are actually compounds of such transposons. This "transposon hypothesis" explains three properties of this repeat family as they were described by Nagaki et al. (Nagaki et al. 1998). First, it was reported that the copy number of *Afa* repeats is highly variable in different Triticeae spp. On one hand, a transposon can be more active in one species than in another and, therefore, produce more copies. On the other hand, we showed that the number of *Afa* repeats can vary drastically even within very closely related

elements, indicating a very rapid evolution of these sequences. Second, the mobility of a transposon explains why no chromosome specificity within one species was observed. Third, Nagaki et al. (Nagaki et al. 1998) suggested the presence of a specific mechanism to remove *Afa* repeats from the genome. The transposon hypothesis can provide this specific mechanism.

The presence of *Afa* and other repeat structures such at *TM-1*, *TM-2*, and the extensive regions comprising short sequence repeats raises new questions. First, the amplification mechanism is still obscure. Template slippage during DNA replication or unequal crossing over can explain the rapid change in copy number, but it does not explain why only some conserved repeat sequences are amplified. A rolling circle amplification, as was suggested by Nagaki et al. (Nagaki et al. 1998), also seems unlikely because it would require a template to be excised from the genome and the amplified product to be reintegrated back into the same element. Second, what is the function of these tandem repeated regions? The presence of such structures in different families of CACTA transposons suggests that they are functional components of these elements rather than the result of random DNA rearrangements.

Despite these open questions, the mere knowledge that tandem repeats are often found within transposons might be important for future analysis of genomic regions. The presence of such arrays can be an indication for the presence of a novel diverged transposon family that could not be detected otherwise. In addition, it is possible that in future studies, tandem repeats from other species such as saccharum CENtromeric sequence repeats from sugarcane (*Saccharum officinarum)* (Nagaki et al. 1998) can be associated with transposons.

The Contribution of CACTA Elements to Genome Evolution

The function and possible benefit of repetitive elements for the "host" plant is a hotly debated question. MITEs, for example, are often found in close association with genes, and they are believed to contribute regulatory sequences that may alter gene expression (Zhang et al. 2000). A similar role can be suggested for CACTA elements. Nine of the total 41 elements were found in EST sequences, suggesting that they may also be found frequently in close proximity to genes. In addition, one *Mandrake* element was found a few kilobase pairs upstream of the *Td-Glu-A3-1* gene in *T. durum* (Wicker et al. 2003b). Interestingly, a different *Mandrake* element was identified at a similar distance to an alpha-gliadin gene in *T. aestivum* (accession no. AF234649). Glutenins and gliadins are genes that belong to the same family. The position of insertion and the

degree of sequence conservation between the two genes indicates that both insertions have been independent events rather than an insertion that occurred already in the common ancestor of the two genes. Therefore, it is possible that certain types of CACTA elements can be involved in specific interactions with certain genes in the Triticeae genomes.

The finding that the four *Caspar* SNAC elements contain sequences similar to 5S rDNA genes is intriguing. The fact that the region that contains the 5S derivative is more conserved among the four elements than the rest of the elements suggests that a selection pressure has been acting on these sequences. It is possible that these sequences have been acquired by a CACTA element during evolution and that they have gained a function that was beneficial for the plant, eventually leading to their fixation within the genome. Acquisition of fragments of cellular genes by CACTA elements has been reported before (Takahashi et al. 1999).

Concluding Remarks

Repetitive DNA, which is still often referred to as "junk DNA," is rarely the focus of a detailed analysis. Our results demonstrate the importance of detailed characterization of repetitive elements and database mining of public databases. Because of their high amount of repetitive DNA, genomic sequences from Triticeae are an essential resource for the identification of novel repetitive elements. The information gained about these elements then can be used for a targeted search for similar elements in other plant genomes. This was demonstrated by the discovery of the rice SNAC transposons, which were not annotated in the publicly available rice sequences. Another important result of our study is the finding that the *CTG-2* protein is actually a part of the *Caspar* transposon. This information suggests that numerous sequences that were interpreted as genes could actually belong to repetitive elements. This has an important implication for future estimates of the total gene contents of entire genomes and also for the calculation of local gene densities in large genome plants such as wheat or maize. Finally, the identification of novel CACTA elements could eventually lead to the discovery of active wheat transposons that could be used for transposon-tagging systems similar to those based on *En/Spm* and *Ac/Ds* elements.

Chapter 5

In silico comparative analysis reveals a mosaic conservation of genes within a novel colinear region in wheat chromosome 1AS and rice chromosome 5S

Romain Guyot, Nabila Yahiaoui, Catherine Feuillet and Beat Keller

(2004)

Functional & Integrative Genomics 4:47-58

5 In silico comparative analysis reveals a mosaic conservation of genes within a novel colinear region in wheat chromosome 1AS and rice chromosome 5S

5.1 Abstract

Comparative RFLP mapping has revealed extensive conservation of marker order in different grass genomes. However, microcolinearity studies at the sequence level have shown rapid genome evolution and many exceptions to colinearity. Most of these studies have focused on a limited size of genomic fragments and the extent of microcolinearity over large distances or across entire genomes remains poorly characterized in grasses. Here, we have investigated the microcolinearity between the rice genome and a total of 1,500 kb from physical BAC contigs on wheat chromosome 1AS. Using ESTs mapped in wheat chromosome bins as additional source of physical data, we have identified 27 conserved orthologous sequences between wheat chromosome 1AS and a region of 1,210 kb located on rice chromosome 5S. Our results extend the orthology described earlier between wheat chromosome group 1S and rice chromosome 5S. Microcolinearity was found to be frequently disrupted by rearrangements which must have occurred after the divergence of wheat and rice. At the *Lr10* orthologous loci, microrearrangements were due to the insertion of mobile elements, but also originated from gene movement, amplification, deletion and inversion. These mechanisms of genome evolution are at the origin of the mosaic conservation observed between the orthologous regions. Finally, *in silico* mapping of wheat genes identified an intragenomic colinearity between fragments from rice chromosome 1L and 5S, suggesting an ancestral segmental duplication in rice.

5.2 Introduction

Earlier studies in comparative genetics performed in the *Poaceae* family revealed that, despite large variations in genome size, there is significant genetic colinearity between grass species that have evolved independently from a common ancestor 50 to 80 million years ago (Wolfe et al. 1989). However, the low resolution of genetic mapping did not allow the detection of sequence rearrangements within apparently colinear regions. The recent increase of available genomic data allowed the comparative analysis of large regions in grass species at the sequence level. Many exceptions to colinearity were reported from such comparisons. In addition to the insertion of transposable elements in the intergenic regions (Tikhonov et al. 1999; Wicker et al. 2001;

SanMiguel et al. 2002; Wicker et al. 2003b), numerous local rearrangements such as gene movement, duplication, deletion and inversion were observed (Dubcovsky et al. 2001; Bennetzen and Ramakrishna 2002; Feuillet and Keller 2002; Song et al. 2002; Brunner et al. 2003). Breaks in microcolinearity indicate that fine-scale comparative analyses remain essential to investigate colinearity and to determine the limits of model genomes for map-based cloning in cereals.

Despite the economic importance of bread wheat, its large genome size (16,000 Mb) (Arumuganathan and Earle 1991) and complex allohexaploid genome organization have limited molecular investigation and sequencing of large genomic fragments in this species. So far, the few sequence data obtained from wheat genomes did not allow comparative studies with rice over large genomic regions. The recent development of large insert bacterial artificial chromosome (BAC) libraries from different wheat species (including diploid, tetraploid and hexaploid wheat species) has promoted the isolation, the physical mapping and the complete sequencing of large genomic fragment (Lijavetzky et al. 1999; Moullet et al. 1999; Ma et al. 2000; Wicker et al. 2001; Cenci et al. 2003; Wicker et al. 2003b). To date, 1.3 Mb of genomic sequences from wheat species with a length >50 kb are available in public databases. In addition, the recent mapping of Triticeae ESTs using wheat deletion lines represents a new source of physical map information for comparative studies (Qi et al. 2003).

In wheat, several genes of agronomical importance have been located in a gene rich region on chromosome group 1S (Gill et al. 1996b; Sandhu et al. 2001). In our group, we have focused our work on chromosome 1AS of wheat and several physical contigs covering three different loci between the RFLP markers *SFR159* and *mwg2245,* anchored on genetic maps, were partially or completely sequenced. At the *Lr10* leaf rust resistance locus, a BAC contig of 450 kb was established in the diploid wheat *Triticum monococcum* (Stein et al. 2000), a model genome for the A-genome of hexaploid wheat (Dubcovsky et al. 1995) and 211 kb were completely sequenced (Wicker et al. 2001). Sequence analysis has revealed the presence of five putative genes, including two disease resistance gene analogs (*RGA1* and *RGA2*), an actin gene (*ACT*), a chromosome condensing factor gene (*CCF*) and a nodulin-like like gene (*NLL*). Furthermore, mapping experiments have shown the presence of a second *NLL* gene, located within the *Lr10* contig, in the overlapping BAC clone 4E14 (Wicker et al. 2001). At the *Lrk10* locus, a genomic fragment of ~16 kb was isolated from *T. aestivum* chromosome 1AS. Two genes which encode different receptor-like kinase proteins (*Lrk10* and *Tak10* genes) as well as a pseudogene similar

to disease resistance gene analogs have been identified (Feuillet and Keller 1999). Recently, two large physical contigs of 470 kb and 180 kb were established at orthologous low-molecular weight glutenin loci (*LMW Glu-A3*), on chromosome $1A^mS$ in the diploid wheat *T. monococcum* and on chromosome 1AS in the tetraploid wheat *T. durum* (Wicker et al. 2003b). In total more than 1,100 kb of physical contigs were generated in the distal part of the wheat chromosome 1AS of which 638 kb of genomic DNA are completely sequenced.

Comparative RFLP mapping experiments between hexaploid wheat and rice have indicated that rice chromosome 5 is largely colinear to the long arm of wheat chromosome group 1 (Ahn et al. 1993; Kurata et al. 1994; VanDeynze et al. 1995a; VanDeynze et al. 1995b). So far, colinearity was only found in the proximal region of the short arm of wheat chromosome group 1 and no relationship could be established in the distal region (VanDeynze et al. 1995b). The availability of the first drafts of two closely related rice genomes (*Oryza sativa* ssp. *japonica* and *Oryza sativa* ssp. *indica,* (Goff et al. 2002), the exceptional source of previous physical mapping and sequencing data on wheat chromosome 1AS (Feuillet and Keller 1999; Stein et al. 2000; Wicker et al. 2001; Wicker et al. 2003b), the establishment of two additional physical contigs at the *bcd1434* and the *SRLK* loci on wheat chromosome 1AS and the physical mapping of a set of wheat ESTs into chromosome bins (Qi et al. 2003) allowed us to investigate microcolinearity between wheat chromosome 1AS and the rice genome over more than one Mb of physical distance. Using computational approaches, we have extended colinearity in the distal regions of wheat chromosome 1S and rice chromosome 5S. This analysis revealed a mosaic structure of conservation between wheat and rice, in which the disruption of colinearity is the rule and not the exception. We also found evidence for an ancient segmental duplication in the rice genome between rice chromosomes 1L and 5S.

5.3 Material and methods

Plant material and genetic mapping in wheat

Genetic mapping of *whs179* and *399A20R* was performed on an F2 population of 1,340 plants derived from a cross between the two *Triticum aestivum* lines, Chul and Frisal. A set of aneuploid nullitetrasomic lines of Chinese Spring (Sears 1966) was also used for mapping in hexaploid wheat. Wheat probe *whs179* was provided by L. Hartl (Freising, Germany) and barley probe

bcd1434 was provided by M. Sorrells (Cornell University, USA). RFLP probes *mwg835* and *mwg938* were provided by A. Graner (IPK Gatersleben, Germany).

BAC library screening, BAC analysis and shotgun sequencing

The screening of the *T. monococcum* BAC library (Lijavetzky et al. 1999), the BAC DNA preparation for fingerprint analysis and BAC end sequencing were performed as described by Stein et al., (Stein et al. 2000). Preparation for low-pass shotgun sequencing of the BAC clones: 539F9, 43J19 and 67C2 (*T. monococcum)* was carried out as described previously by Stein et al., (Stein et al. 2000).

Southern analysis

The following *Triticum aestivum* lines were used for mapping in chromosome 1AS: Chinese Spring, the Chinese Spring nullitetrasomic line N1A/T1B and five Chinese Spring deletion lines of chromosome 1A produced by Endo and Gill (Endo and Gill 1996). The order of breakpoints of these lines was as follows: centromere (FL value 0.00), 1AS-5 (FL value 0.20), 1AS-2 (FL value 0.45), 1AS-1 (FL value 0.47), 1AS-4 (FL value 0.76), 1AS-3 (FL value 0.86), and telomere (FL value 1.00). FL refers to fraction length of the remaining part of the chromosome 1AS arm. Information about deletion lines is available at http://www.ksu.edu/wgrc/Germplasm/Deletions/group1.html. Procedures used for genomic DNA isolation, restriction endonuclease digestion, gel electrophoresis and DNA gel blot hybridization were described in Stein et al., (Stein et al. 2000). The three probes *F640* (*Actin,* (Stein et al. 2000)), *Lrk10* (Feuillet and Keller 1999) and *SFR159* (Wicker et al. 2001) were used for Southern blot analysis.

Wheat and rice sequence data and map sources

The sequences of RFLP probes located on the genetic map of chromosome group 1S and used for comparative mapping among grasses were downloaded from the NCBI web site (http://www.ncbi.nlm.nih.gov/). These RFLP probes and additional specific genomic markers for chromosome 1AS developed and sequenced by our group were used for BLASTN searches (Altschul et al. 1997). A consensus genetic map for wheat chromosome 1S was drawn with RFLP probes which have been sequenced and were available in databases. The order of these RFLP

probes along chromosome 1S was deduced from consensus genetic linkage maps (summarized in (Sandhu and Gill 2002b)). Other wheat genomic sequences were downloaded from GENBANK (*Lr10* locus: AF326781; *Lrk10* locus: U51330, U78762 and U76215; *LMW Glu-A3* locus: AY146588 and AY146587), or came from low-pass and BAC end sequencing (539F9F : CG892533; 67C2R207 : CG876950;. 67C2F260 : CG876951; 399A20R : CG876949). The sequences of *O. sativa* ssp. *japonica* BACs were downloaded from the RiceGAAS web site (ftp://ftp.dna.affrc.go.jp/pub/RiceGAAS), as of May 2003. The database has a size of 516 Mb included a total of 3,782 BAC sequences for which 32% of the sequences are completely finished. The positions of BACs within the rice genome were estimated using the FPC (Finger Printed Contigs) physical map available at http://www.genome.arizona.edu/fpc/rice/ and using the IRGSP mapping data (http://rgp.dna.affrc.go.jp/IRGSP). Triticeae EST sequences anchored within chromosome 1AS and assigned by chromosome bin mapping were downloaded from the GrainGenes web site http://wheat.pw.usda.gov/NSF/progress_mapping.html.

Physical map of the distal region of rice chromosome 5S

BAC assembly in the distal region of rice chromosome 5S was performed using data from the FPC rice physical map. In order to identify and confirm candidate overlapping clones, BAC sequences from FPC contigs 102 and 103 of chromosome 5S were first used for a BLASTN search against rice genomic sequences. Quality and length of overlaps were checked by pair using alignments displayed by dot plot (DOTTER, (Sonnhammer and Durbin 1995)). BAC clones were assigned as members of the physical contig created on chromosome 5 if they had at both ends a 100% identity at the nucleotide sequence over a minimum length of five kb with other BAC clones belonging to the same FPC contig. At one location within the created contig on rice chromosome 5S, a BAC clone from *O. sativa* ssp. *indica* (AF532975) was used to remove ambiguities and to confirm overlaps between the two following adjacent BACs clones: AC104285 and AC079022. The *indica* BAC clone AF532975 displayed at one end a nucleotide identity of >99% over a distance >29 kb with the BAC clone AC104285 and at the other end an identity of >99% over a distance >24 kb with the BAC clone AC079022. By including the *indica* BAC clone within the contig, an uninterrupted physical contig of ~1,420 kb was created in *O. sativa* ssp. *japonica* assembling 14 BAC clones (for 1,951 kb of sequences) within the FPC contigs 102 and 103. Similarly, a contig of 9 BAC clones was assembled in a fragment of ~1,130

kb in the FPC contig 27 of rice chromosome 1 (for 1,435 kb of sequences). GenBank accessions of rice BACs given in Figures 3, 4 and 5: chromosome 1 AP004326, AP003238, AP003263, AP004365, AP003448, AP003277, AP003298, AP003627; chromosome 5 AC073405, AC084818, AC104285, AC129716, AC079022, AC079356, AC09321, AC108504, AC078977, AP001111, AC093088, AC093089, AC079021, AC134931.

BLAST analysis

To identify rice BACs with nucleotide sequence similarity to the wheat sequence queries, we used BLASTN searches performed against a local database composed of all plant sequences (containing non-redundant nucleotide, EST, GSS, and HTGS sequences) and using a local BLAST server. Following BLAST analysis, results were parsed to eliminate low identity sequences with a threshold limit of score > 50. Results were sorted according to the score, the E-value, the rice BAC accession, the chromosome location and the coordinates within the rice BACs. Redundancies were removed from raw data, according to the following parameters: (i) if a wheat sequence gave several hits in the same rice BAC sequence, only the highest hit was kept; (ii) if a wheat sequence detected two hits with strictly identical scores and E-values within adjacent rice BAC clones, a single rice BAC was reported.

Sequence analysis

DNA sequences were analysed using BLASTN algorithms against public and local nucleotide databases. Detailed analysis was performed with the GCG Sequence Analysis Software Package version 10.1 (Madison, WI), the EMBOSS package (Rice et al. 2000) and by dot plot (DOTTER). Annotation of "low-pass" sequences (*SRLK* locus) was performed by similarity search.

5.4 Results

Extension of a partial physical map in a distal region of chromosome 1AS in wheat

Three loci (*Lr10*, *Lrk10* and *LMW Glu-A3*) have been previously analysed and completely sequenced on chromosomes $1A^{m}S$ in *T. monococcum* as well as 1AS in *T. durum* and *T. aestivum* (Stein et al. 2000; Feuillet et al. 2001; Wicker et al. 2001; Wicker et al. 2003b) (Figure 1). To

extend physical mapping, two additional regions have been investigated in this sub-telomeric region. A large contig of 335 kb was established at the *SRLK* locus in *T. monococcum*. This region which is anchored to the wheat genetic map by the three RFLP probes *mwg938, mwg835* and *whs179* (Figure 1) was low-pass sequenced. The analysis of shotgun and BAC end sequences by similarity search revealed the presence of three putative genes: a putative receptor protein kinase (*SRLK*/539F9F CG892533) a disease resistance gene analog (*NBS-LRR/*67C2F260 CG876951) and a gene similar to a hypothetical protein in rice (*Unknown* protein/67C2R207 CG876950). This BAC contig is located between the proximal *Lrk10* and the distal *LMW Glu-A3* loci (Figure 1). A second region composed of the BAC clone 399A20 was located at the *bcd1434* locus in *T. monococcum* and a BAC end probe (*399A20R,* CG876949) was mapped (Figure 1). In total, five regions were analysed in the distal part of wheat chromosome 1AS, representing more than 1,500 kb of physical contigs.

Identification of probes conserved between wheat chromosome group 1S and rice BAC sequences

In order to identify putative colinear regions between wheat chromosome group 1S and the draft sequence of the rice genome (*Oryza sativa* ssp. *japonica*), we have used the nucleotide sequences of 18 RFLP probes as well as two sequences corresponding to the probes *SFR159* (Wicker et al. 2003b) and *whs179* (unpublished data) to compare with the rice genomic BAC sequences. Thirteen probes (60%) showed significant similarity to rice (Supplement 1). The four probes *bcd1434*, *cdo580*, *rz244* and *cdo1173* identified rice BACs located on three adjacent FPC (Finger Printed Contigs) physical contigs (contigs 102, 103 and 104) at the top of chromosome 5 (Figure 2). The remaining probes showed a range of hits distributed on 10 of the 12 chromosomes of the rice genome. These data confirm previous comparative linkage analysis, where the three markers *cdo580*, *cdo1173* and *rz244* were mapped in the proximal part of the rice chromosome 5S and on the short arm of wheat chromosome group 1 (Ahn et al. 1993; Kurata et al. 1994; VanDeynze et al. 1995a; VanDeynze et al. 1995b). Due to a lack of polymorphism, *bcd1434* was not mapped in rice (VanDeynze et al. 1995b). However, its *in silico* detection together with the markers *cdo580*, *cdo1173* and *rz244* in a region colinear between wheat and rice indicates a putative extension of colinearity between the short arms of orthologous rice chromosome 5 and wheat chromosome 1.

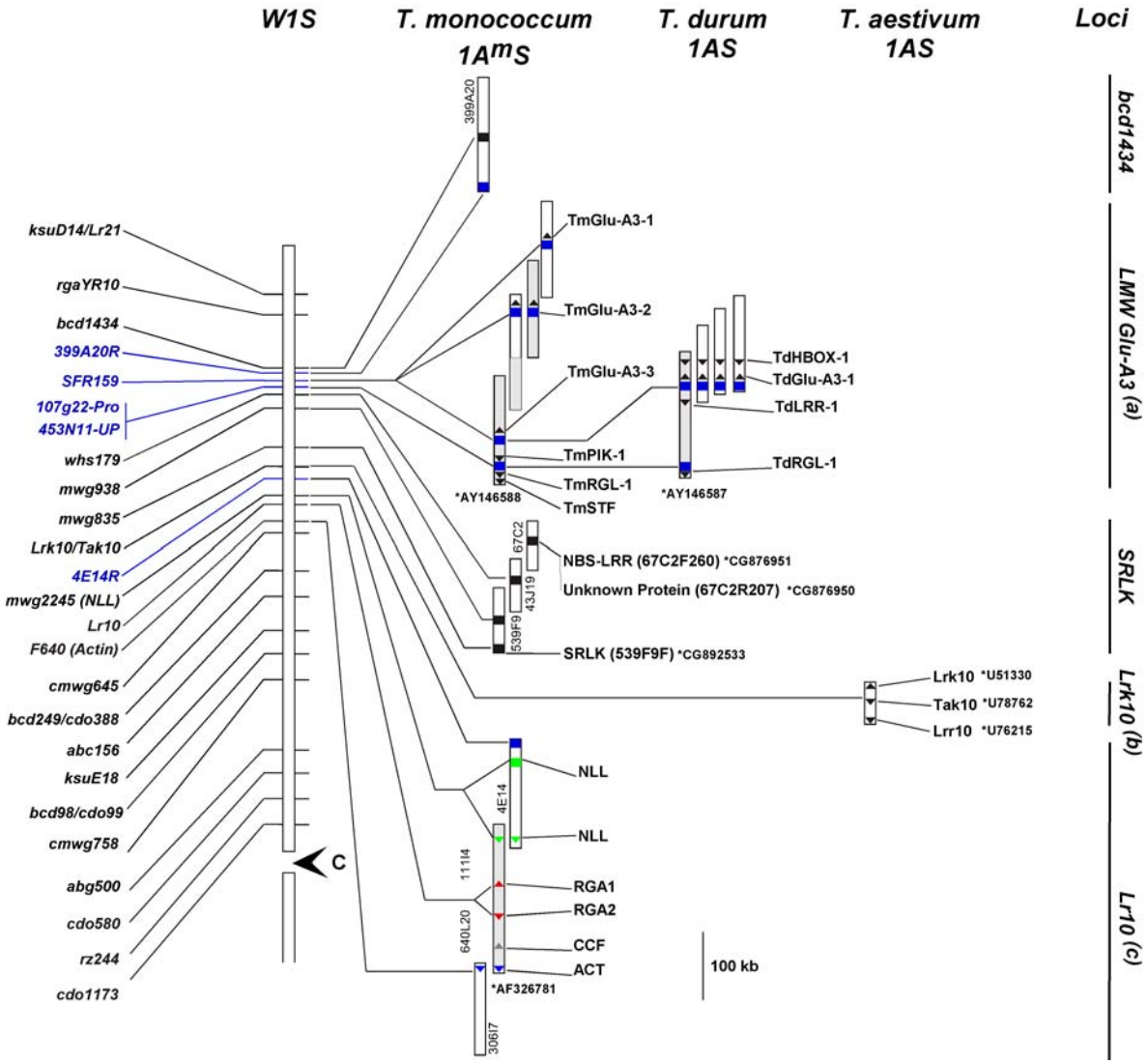

Figure 1 Schematic representation of five physical loci located along the chromosome 1AS of three different wheat species. The genetic map positions of RFLP markers derived from cDNAs of wheat chromosome group 1S and five genomic sequences (in blue) on chromosome 1AS are indicated on the left side of the figure. Contigs are represented by boxes and are sorted according to the wheat species (*T. monococcum*, *T. durum* and *T. aestivum*). Grey contigs: completely sequenced BAC clones and white boxes: BAC clones anchored within the genetic map. The physical position of genes is indicated by arrowheads and the direction of arrowheads indicates the transcriptional orientation of genes. Relative positions of markers and genes identified by low-pass sequencing are indicated by coloured squares. The names of genes within the contigs are reported as indicated in Supplement 3. (*) indicate the accession numbers of sequences. References: (a): *LMW Glu-A3,* Wicker et al., 2003; (b): *Lrk10*, Feuillet and Keller, 1999; (c): *Lr10*, Stein et al., 2000.

Identification of rice sequences homologous to genes located on the wheat chromosome 1AS physical contigs

To study in more detail the colinearity relationships between wheat chromosome 1S and rice chromosome 5S, the nucleotide sequences of all 20 sequences predicted as genes, pseudogenes or part of genes on the wheat physical contigs (Figure 1) were used as queries to perform a BLASTN search against the *O. sativa* ssp. *japonica* genomic BAC sequences. Only two genes, *ACT* and *CCF,* of the *Lr10* locus gave a hit on chromosome 5 (Supplement 2). Both genes identified orthologs on the FPC contig 102 (BAC clone AC104285), which was previously identified with *bcd1434*. In addition, both genes identified a common contig in a distal region of the long arm of rice chromosome 1 (FPC contig 27). These data indicate that the *ACT* and *CCF* genes are closely linked and are conserved on two different chromosomes in the rice genome. The BLAST analysis of the *NLL* gene of the *Lr10* locus revealed a lower score than the threshold applied for the search. A more sensitive TBLASTX search was then used to test for the presence of the *NLL* gene in rice. BLAST search detected two *NLL* genes tandemly repeated on the BAC clone AC104285 on chromosome 5 (FPC contig 102) where the *ACT* and *CCF* homologous genes are located, but not on chromosome 1 FPC contig 27. In addition, BLAST search with wheat genes as queries identified a BAC clone in *O. sativa* ssp. *indica* (AF532975) carrying three genes showing homology to the *ACT*, the *CCF* and the *NLL* genes in wheat. The overall nucleotide identity >99% over a distance of ~29 kb between the *indica* BAC AF532975 and the *japonica* BAC AC104285 indicates orthology between these two genomic fragments.

Thus, our data indicate that four genes located at the wheat *Lr10* locus (*ACT*, *CCF* and two *NLL)* are conserved and located on the same BAC clone in the rice *japonica* subspecies. These data support the hypothesis of a partial orthology between wheat chromosome 1AS and the short arm of rice chromosome 5. Furthermore, the similarities found with the distal part of rice chromosome 1L suggest a duplication between chromosome 5S and 1L in rice.

Extension of wheat-rice colinearity by identification of conservation between Triticeae ESTs anchored in wheat chromosome 1AS bins and the rice genome

So far, only a limited set of sequences was available from wheat chromosome 1AS to investigate colinearity. The recent assignment of Triticeae ESTs to chromosome bins in wheat (Qi et al. 2003) provided a new source of sequences with physical mapping data for comparison and

validation of colinearity with rice chromosome 5. Three bins have been assigned along chromosome 1AS of wheat.

Sixty-three of the 101 ESTs sequences (62%) from the distal bin 1AS3-0.86-1.00 showed significant similarity to one or more genomic clones in rice (Supplement 3). A similar ratio (54/88, 61.3%) was obtained with the middle bin 1AS1-0.47-0.86. The nine ESTs of the proximal bin C-1AS1-0.47 were all conserved in rice (data not shown). In all cases, a bias to chromosome 5 sequences was observed (Figure 3A). For the distal bin, 21 ESTs sequences showed similarities with the BAC clones belonging to the FPC contigs 102 and 103 on chromosome 5. Considering the relative size of rice chromosomes and the non-availability of the complete sequence of the rice genome, the remaining EST sequences seemed to have a random distribution over the 12 rice chromosomes. For the middle bin, 20 EST sequences showed similarities with sequences of the BAC clones belonging to FPC contigs 103/104 of chromosome 5 (Figure 3A). For the proximal bin, only three ESTs showed similarities to sequences located on chromosome 5 in three different FPC contigs (103, 104 and 105). The examination of all the BLAST hits indicated the presence of an additional group of homology with the ESTs on rice chromosome 1 (Supplement 3). Three ESTs which have been located on the rice chromosome 5S colinear region (Supplement 3, ESTs 6, 8, 11) and an additional EST (WHE1071-1074_H24_H24ZS, E-value = 6E-42) were found in the FPC contig 27 of rice chromosome 1L. Together with the *ACT* and the *CCF* genes, a total of six genes located on chromosome 1AS of wheat was found to be conserved in a contiguous segment of ~660 kb of the 1,130 kb of contig 27 of rice chromosome 1L. The redundancy and the conservation of the order of the ESTs located in the rice genome suggest intragenomic colinearity, involving regions from rice chromosomes 1 and 5.

Our results indicate colinearity between the RFLP marker *bcd1434*, the *ACT*, the *CCF*, and the *NLL* genes in wheat and the FPC contig 102 in rice, as well as between the wheat chromosome bin 1AS3-0.86-1.00 (six ESTs conserved) and the rice FPC contig 102 (Figure 3A). This led us to investigate the physical location of the markers from the wheat physical contigs in the wheat chromosome 1AS bins (Figure 3A). Two sequences belonging to the distal part of wheat chromosome 1AS (*Glu-A3*, and *TdLRR*) were found *in silico* in the bin 1AS3-0.86-1.00, whereas the marker *rz244*, which is located in the proximal part of chromosome 1AS, was detected in the adjacent bin 1AS1-0.47-0.86. No markers or genes were detected *in silico* in the proximal bin C-1AS1-0.47. Because few sequences located within our physical contigs were found *in silico* in

the wheat chromosome 1AS bins by sequence comparison, we have determined the bin location of three RFLP markers *SFR159* (*LMW Glu-A3* locus), *Lrk10* (*Lrk10* locus) and *F640* (Actin gene: *ACT*, *Lr10* locus) by Southern hybridization experiments on Chinese Spring deletion lines for chromosome 1AS (Endo and Gill 1996). All three markers mapped to the first distal deletion of chromosome 1AS between the FL value 0.86 and 1.00 (data not shown) confirming that the bin 1AS3-0.86-1.00 which is colinear with the FPC contigs 102 and 103 of rice, carried all the wheat BAC contigs in this study.

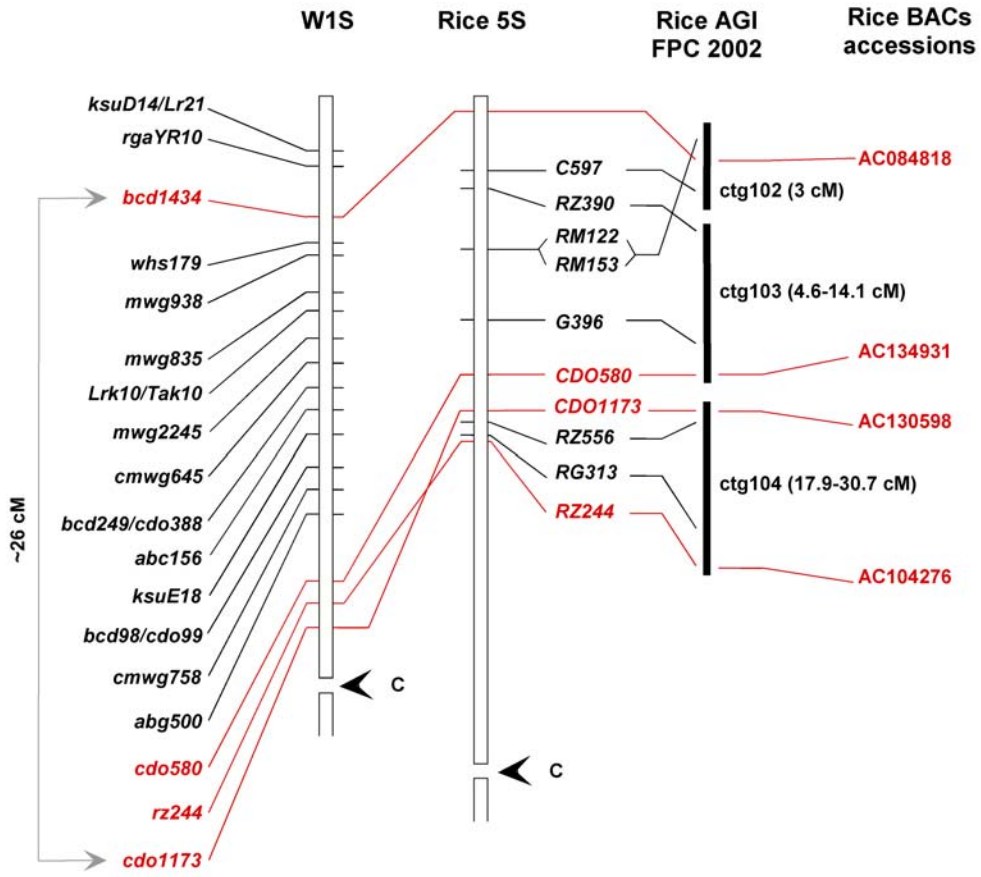

Figure 2 Comparative map of RFLP markers from the short arm of wheat chromosome group 1, the short arm of rice chromosome 5 and a partial physical map of three adjacent FPC contigs of rice chromosome 5S. The comparative map between the rice genetic map and the FPC (Finger Printed Contig) physical contigs was drawn according to the data available at the Gramene (http://www.gramene.org/) and at the Arizona Genomics Institute web sites (http://www.genome.arizona.edu/fpc). Recombinational distances in wheat are taken from a genetic linkage map (Sandhu and Gill, 2002). Accession numbers of rice BAC clones carrying homologs of wheat RFLP markers are indicated. Markers that are conserved between wheat and rice are in red.

Dot plot alignments were performed between the 21 Triticeae ESTs of the proximal bin (1AS3-0.86-1.00) which are conserved in rice chromosome 5S and the rice BAC sequence of the FPC

contigs 102 and 103. This showed that rice homologs are distributed along the contigs on chromosome 5S (Figure 3B). Three Triticeae ESTs (4, 5, 6, Supplement 3) were found within the rice BAC clone (AC104285) carrying homologs of the wheat *ACT*, *CCF* and *NLL* genes. In total, 27 homologous sequences (the *bcd1434* and *cdo580* probes, the *ACT*, *CCF* and 2 *NLL* genes and 21 Triticeae ESTs) of chromosome group 1S in wheat were identified in a genetic region which covers ~26 cM in wheat (between probes *bcd1434* and *cdo580,* (Sandhu and Gill 2002b)) and a physical distance of 1,210 kb in rice (FPC contigs 102 and 103, Figure 3B). These data suggest a limited collinear conservation between wheat and rice of about 20% of all the sequences tested.

Analysis and detailed sequence comparisons of the Lr10 homologous regions in wheat and rice

Detailed comparative analysis was performed at the sequence level between colinear regions of the wheat *Lr10* locus (Wicker et al. 2001), the *O. sativa* ssp. *japonica* (*Osj*) BAC clone AC104285 of chromosome 5, its putative ortholog in the *O. sativa* ssp. *indica* (*Osi*) BAC clone AF532975 and an *Osj* BAC contig of chromosome 1 (AP003263 and AP004365). Pairwise comparisons, using the program DOTTER revealed the presence of five different conserved segments across the four regions studied. Conservation was limited to the exons of three gene families: *ACT*, *CCF* and *NLL*. Two *NLL* genes are present both in wheat and in the *japonica* chromosome 5, while the sequenced BAC in *indica* covered only one gene which is located at one end of the BAC (Figure 4). No *NLL* gene was observed in contig 27 of chromosome 1L in rice.

The *ACT* and *CCF* genes were conserved in the four regions with different relative orientations. While the two genes are in opposite transcriptional orientation in all the regions studied, the wheat *CCF* gene is proximal to the *NLL* gene whereas in the *japonica* chromosome 5 and in its orthologous *indica* fragment, the *CCF* gene was distal to the *NLL* gene. In addition, conservation of the gene order between rice regions in chromosomes 1L and 5S indicates an inversion of the *CCF* and *ACT* genes in rice chromosome 1L compared to wheat.

Intergenic regions between the *NLL* and the *CCF* genes showed a large expansion of size in wheat compared to the orthologous regions in *japonica* and *indica* rice genomic regions. In wheat, the distance between the genes is 178 kb while in the two rice subspecies the *NLL* and the *ACT* genes are separated by 7 kb. This represents an expansion of distance by a factor of >25.

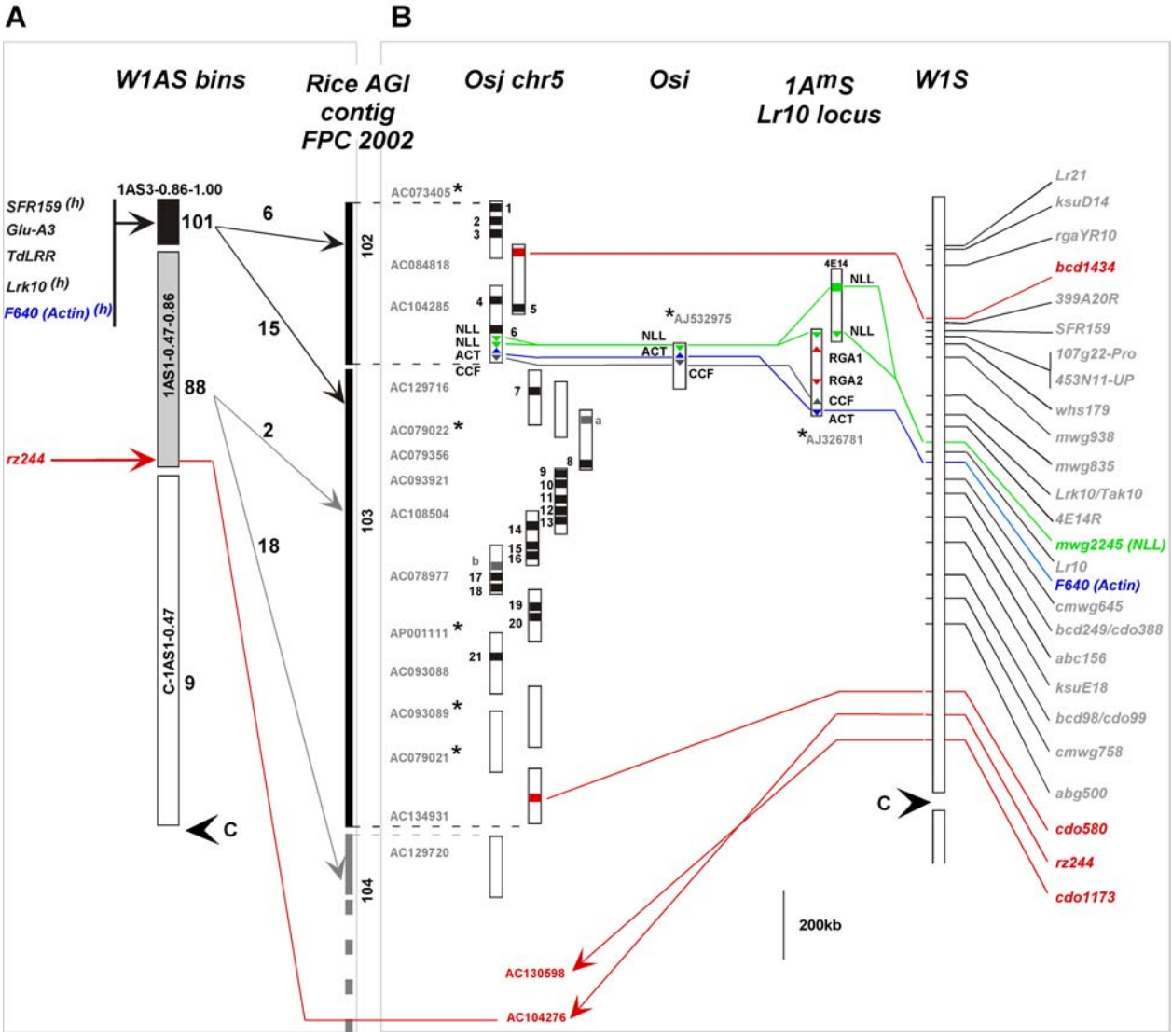

Figure 3 Comparative maps of sequences conserved between the genetic map of wheat chromosome group 1, the physical map of the wheat *Lr10* locus, the partial physical map of the distal part of rice chromosome 5S and ESTs mapped within chromosome 1AS bins of wheat deletion lines.

A. Representation of ESTs conserved within chromosome 1AS bins and FPC contigs 102, 103 and 104 in the short arm of rice chromosome 5. The markers mapped in chromosome bins by BLAST analysis and by hybridization experiments (h) carried out with Chinese Spring deletion lines are indicated at the left side.

B. Detailed comparative representation of conserved Triticeae ESTs of bin 1AS3-0.86-1.00 with the FPC contigs 102 and 103 in *O. sativa* ssp. *japonica (Osj)*, the BAC clone AJ532975 in *O. sativa* ssp. *indica (Osi)*, the wheat *Lr10* locus and the genetic map of chromosome 1S of wheat. Black boxes indicate the presence of conserved sequences between ESTs from bin 1AS3-0.86-1.00 and rice contigs. Numbers indicate the conserved EST sorted as in Supplement 2. Grey boxes indicate conserved sequences with two ESTs from bin 1AS1-0.47-0.86 (a: WHE0333_G09_M17zs, b: WHE-0990_M13_M1). Coloured arrowheads indicate location and orientation of genes as follows: red: *RGA* (resistance genes analogs) genes; grey: *CCF* (chromosome condensing factor) genes; blue: *ACT* (actin) genes; and green: *NLL* (nodulin–like like) genes. Red squares within rice BAC clones refer to genetically mapped wheat RFLP probes conserved in rice genomic contigs (see also Figure 2). Stars associated to accession numbers indicate finished sequences of wheat and rice BAC clones.

In wheat, in addition to the insertion of numerous mobile elements (Wicker et al. 2001), the non conserved interval contains additional genes (Figure 4). First, two Resistance Gene Analogs (*RGA1* and *RGA2*) are present in wheat, but absent in rice colinear contigs. However, homologs of these genes were identified in the rice genome at non orthologous locations (Supplement 2). In addition, a putative gene was predicted in the rice interval (data not shown). In wheat, dot plot analysis has also revealed the presence of an additional gene within the interval between the *NLL* and the *CCF* genes, which has not been discovered earlier (Wicker et al. 2001). The novel gene called $CCF^{(p)}$ showed a nucleotide conservation with the two last exons of the *CCF* gene in wheat. However, no clear ORF was observed, suggesting a partial and non functional copy of the *CCF* gene. The two genes separated by 2 kb of sequence are tandemly repeated. These data suggest an ancient duplication event and rearrangements at the origin of the presence of the $CCF^{(p)}$ gene.

The intergenic regions between the *ACT* and the *CCF* genes also showed major alterations. The distance between the wheat genes (15 kb) is two fold larger than the distance between the orthologous genes on *japonica* chromosome 5 and on *indica* (6.3 and 7.5 kb respectively), but >3 fold lower than the interval between homologous genes on *japonica* chromosome 1 (56 kb, Figure 4). In wheat, two partial retroelements (Wicker et al. 2001) contributed to these alterations, whereas in rice the variation observed between the *japonica* chromosome 5 sequence and the *indica* sequence is due to the insertion of a 1,249 bp of sequence in *indica*. This insertion shows similarities of 65 bp at each end with the sequence of MITE-adh, type M (Tarchini et al. 2000). In the rice chromosome 1 homologous region, the large size increase between the *ACT* and the *CCF* genes was attributed to the presence of a stretch of more than 42 kb of retroelements. These data suggest that multiple local rearrangements involving genes and repetitive elements have contributed to the disruption of colinearity observed between wheat and rice and to the different expansion of the length of paralogous segments in rice.

Analysis of the large duplication between the rice chromosomes 1 and 5

We have shown an intragenomic conservation of genes between the rice contigs 102 and 103 on chromosome 5S and the rice contig 27 on chromosome 1L. All the BAC sequences from the two chromosomal regions were aligned with the DOTTER program. Figure 5A shows a dot plot covering two partial regions of the BAC clone AC079022 in contig 103 of chromosome 5 and the

BAC clone AP003277 in contig 27 of chromosome 1. Six conserved regions, mostly limited to annotated genes were identified among ~93 kb on chromosome 5 and 40 kb on chromosome 1 (Figure 5A). Duplicated genes were separated by variable stretches of completely different regions including annotated genes not retained by the duplication (Figure 5).

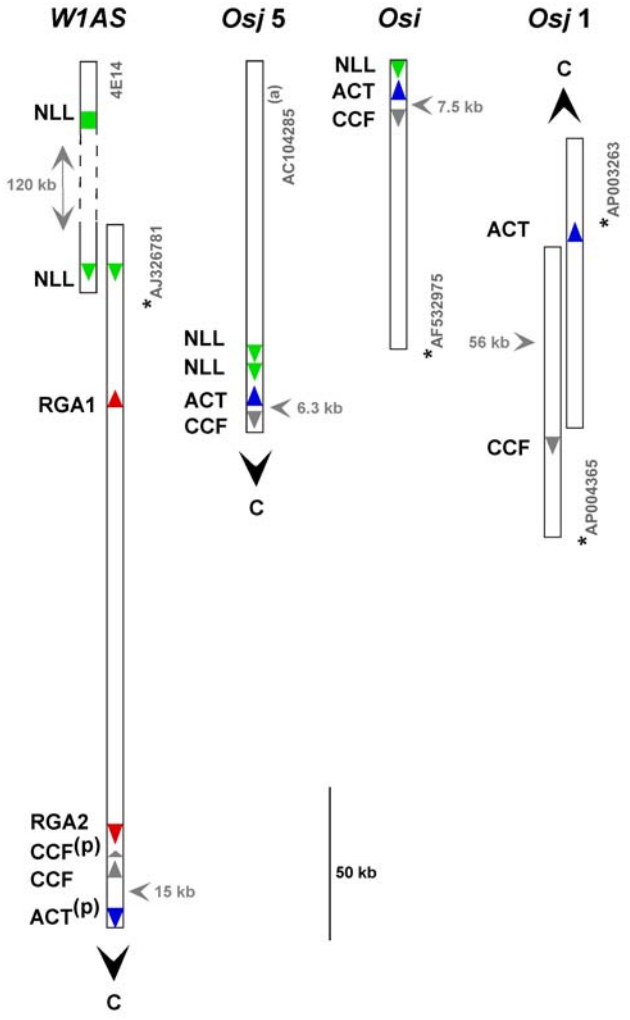

Figure 4 Comparative structural analysis of the *Lr10* homologous regions from wheat (*T. monococcum,* W1AS), rice ssp. *japonica* chromosomes 5 (*Osj* 5) and 1 (*Osj* 1) and rice ssp. *indica* (*Osi*). Coloured arrowheads indicate location and transcriptional orientation of genes as follows: red: *RGA* (resistance genes analogs) genes; grey: *CCF* (chromosome condensing factor) genes; blue: *ACT* (actin) genes; and green: *NLL* (nodulin–like like) genes. The positions of centromeres (C) are indicated. The distance in kb between *CCF* and *ACT* genes is also indicated in each contig. [(a)] Sequence of the BAC clone in 6 ordered pieces. [(p)] Partial genes.

The two duplicated segments are located in the same orientation at the ends of the long arm of rice chromosome 1 and the short arm of rice chromosome 5. At the sequence level, the order and the orientation of the duplicated annotated genes is conserved, except for an inversion involving

gene P0459B04.9 (chromosome 1 BAC AP003627, Figure 5). In total, 22 annotated genes are conserved between a segment of ~665 kb on contigs 102 and 103 of rice chromosome 5 and ~870 kb on contig 27 of rice chromosome 1 (Figure 5B). These data indicate that these two regions originate from a segmental duplication in rice between chromosomes 1 and 5.

5.5 Discussion

Extension of colinearity between the short arms of wheat chromosome 1 and rice chromosome 5

Earlier comparative mapping studies have established colinearity between the short arms of wheat chromosome 1 and rice chromosome 5. So far the colinearity was limited to proximal regions (Ahn et al. 1993; Kurata et al. 1994; VanDeynze et al. 1995a; VanDeynze et al. 1995b). Here, *in silico* analysis extended this colinearity to the distal parts of the wheat chromosome 1AS and rice chromosome 5S. The order of the RFLP probes, the relative order of wheat chromosome bins containing homologs of rice chromosome 5S genes as well as the conservation of several genes at the *Lr10* locus in wheat suggest a colinearity involving the whole short arms of wheat chromosome 1A and rice chromosome 5. This result demonstrates that computational analysis can greatly help to detect additional colinear relationships between distantly related species in the grass family. However, such an analysis is limited by the availability of large genomic sequences in other species than rice. The use of a large set of ESTs mapped within chromosome 1AS bins of wheat deletion lines (Qi et al. 2003) proved to be a useful tool to validate and complete data obtained by genetic mapping and sequencing of large stretches of genomic DNA. Similarly, a recent analysis based on a comparison between 4,485 physically mapped wheat ESTs and the ordered part of the rice genome has shown the usefulness of such approaches to investigate colinearity at the genome level (Sorrells et al. 2003a).

In the region studied here, the degree of the colinearity is very low. From the 1,500 kb of physical contigs and the 638 kb sequenced on wheat chromosome 1AS, on a total of 20 sequences predicted as coding regions only four were colinear with the rice chromosome 5S. In total, with RFLP markers from wheat chromosome group 1S and ESTs from wheat chromosome 1AS bins, less than 20% of the tested sequences are retained in colinearity. Similar observations were made at the genome level by comparative sequence analysis between wheat ESTs and the

rice genome (Sorrells et al. 2003a) and by a statistical reanalysis of comparative mapping data (Gaut 2002). These data suggest that the grass genomes are less conserved than previously reported.

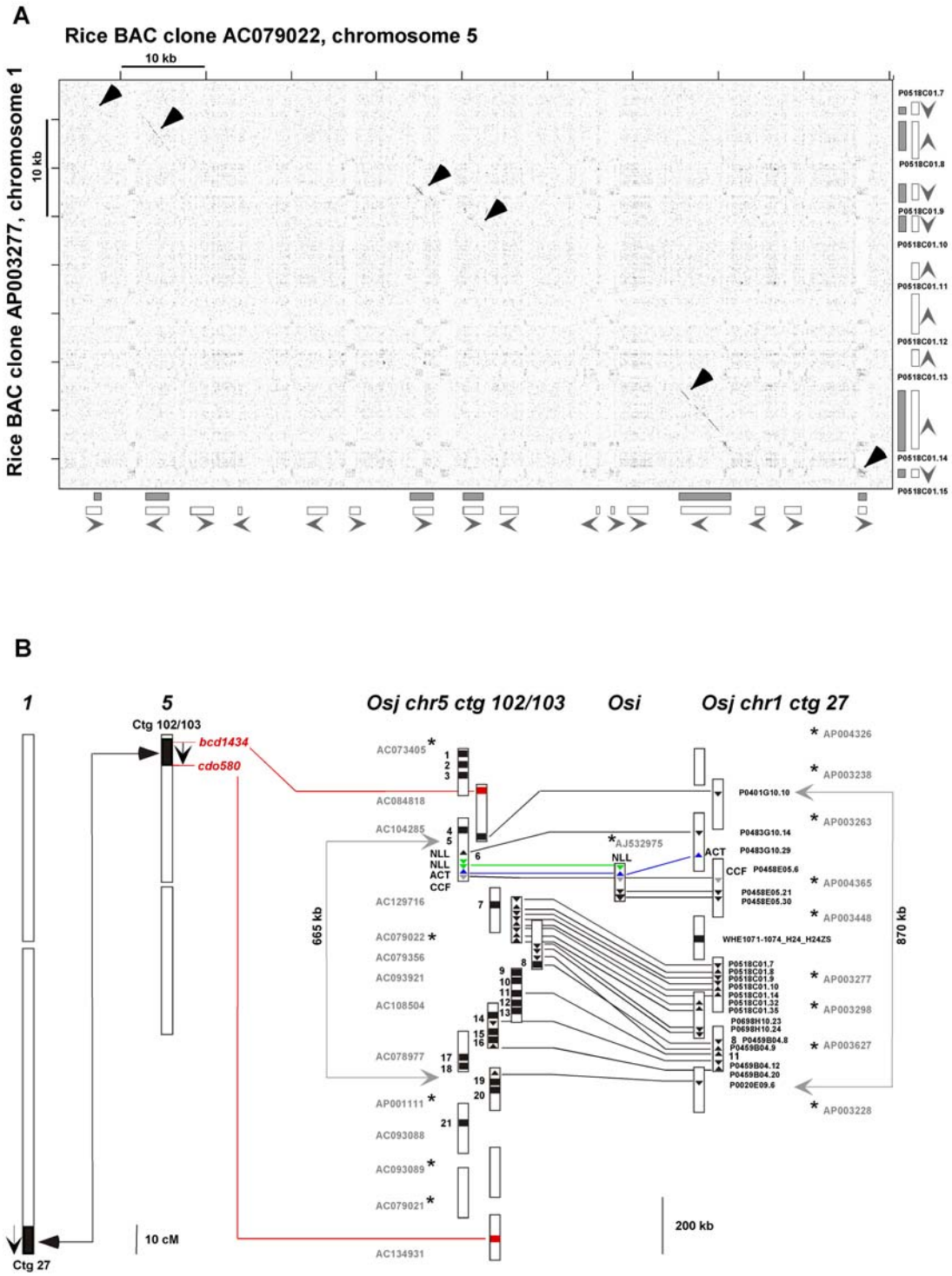

Figure 5 Analysis of the large duplication involving rice chromosome 1 contig 27 and chromosome 5 contigs 102 and 103

A. Dot plot alignment between the partial sequences of the BAC clone AC079022 (horizontal axis) from rice chromosome 5 and the BAC clone AP003277 (vertical axis) from rice chromosome 1. Black arrowheads indicate the positions of duplicated genes. For the two BAC clones, annotated genes are represented by white boxes and conserved sequences are represented by grey boxes. Grey arrowheads represent the predicted transcriptional orientation of genes. The accession numbers of annotated genes are given for the chromosome 1 BAC AP003277. For the chromosome 5 BAC AC079022, gene predictions are given by Rice Genome Annotation Database (RiceGAAS).
B. Schematic representation of the duplicated regions between the rice chromosomes 1 and 5 from *japonica* and the BAC clone AJ532975 from *indica*, orthologous to AC104285 (*OSj* chr5). Location and orientation of the duplication are represented respectively by black boxes and black arrows in the schematic representation of the two rice chromosomes at the left. The Triticeae ESTs conserved between the wheat chromosome 1AS bin (1AS3-0.86-1.00) and the contig 27 (chromosome 1) and contig 102/103 (chromosome 5) are indicated by black boxes and numbers (reported in Supplement 2). Black arrowheads give the location and the orientation of genes conserved between the rice chromosomes 1 and 5. When annotations of BAC clones are available in GENBANK, the accession numbers of the conserved genes are given. Coloured arrowheads indicate the location and the orientation of genes as follows: grey: *CCF* (chromosome condensing factor) genes; blue: *ACT* (actin) genes; and green: *NLL* (nodulin–like like) genes. Red squares within rice BAC clones refer to genetically mapped wheat RFLP probes conserved in rice contigs (see also Figure 2). Stars associated to accession numbers indicate finished sequences of BAC clones.

Mosaic pattern of conservation at orthologous loci between wheat and rice

Our results indicate that between wheat chromosome 1AS and rice chromosome 5S a low degree of colinearity is maintained within a mosaic structure composed of islands of conserved sequences, separated by large stretches of non conserved regions. This complex and restricted pattern of conservation raises the question about the molecular mechanisms which are responsible for the dynamic evolution of the grass genomes.

Disruption of colinearity between wheat and rice was observed here at two different levels. On wheat chromosome 1AS, three gene-rich loci: *LMW Glu-A3*, *SRLK* and *Lrk10*, covering at least 880 kb of sequence were not found on rice chromosome 5S. These loci contain resistance gene analogs (NBS-LRR) and storage protein genes (LMW glutenin genes). Previous mapping experiments at resistance genes loci have revealed a limited orthology across grass species (Gallego et al. 1998; Leister et al. 1998). Homologs of resistance genes located within the *LMW Glu-A3* were found conserved at non orthologous positions on rice chromosomes 1, 2, 3, and 10. Homologs of *Lrk10* and *Tak10* receptor-like kinase genes (*Lrk10* locus) were present in a cluster of at least ten members at a non colinear location in rice chromosome 1S (Keller and Feuillet 2000). The rice chromosome 1S is homologous to chromosome group 3 in Triticeae and chromosome 8 in maize in which *Lrk* and *Tak* homologous genes have also been found (Feuillet and Keller 1999). In addition, the rice genome does not contain *LMW glutenin* genes homologs. These storage protein genes have probably arisen in wheat after the separation of the two species,

participating also to the large disruption of colinearity observed in this region between wheat chromosome 1AS and rice chromosome 5S. Both wheat storage protein and disease resistance genes form large gene families resulting from gene amplification events ((Wicker et al. 2003b) and unpublished data) that probably contributed to the large size of the non conserved region. A mosaic organization of orthologous sequences was first described between more than 400 kb of maize, sorghum and rice sequences corresponding to the maize 22-kD zein cluster (Song et al. 2002). A similar structure of conservation and non conservation observed here at a large scale between wheat and rice confirms that gene movement and amplification are a major factor of the evolution of the grass genomes.

In addition to this large disruption of colinearity, comparative sequence analysis revealed a number of local rearrangements within colinear *Lr10* loci. Large variations in the size of intergenic regions due to differential insertion of repetitive elements were found. As was observed previously in grass genomes (Chen et al. 1997; Chen et al. 1998; Tikhonov et al. 1999; Tarchini et al. 2000; Dubcovsky et al. 2001; Song et al. 2002; Brunner et al. 2003; Klein et al. 2003; Wicker et al. 2003b), expansion and contraction of the genome due to insertions and deletions of transposable elements promote a fast evolution of grass genomes and contribute to the observed disruption of the colinearity between wheat and rice. Finally, significant rearrangements involving genes were found. These included a duplication of the *CCF* gene in wheat compared to rice, a translocation of the *RGA1* and *RGA2* genes in rice compared to wheat and a local inversion of the *ACT* and *CCF* genes in rice compared to wheat. Similar disruptions in partially colinear loci have been reported by comparative analysis at the *adh1* and the *zein* gene regions between rice, sorghum and maize (Tikhonov et al. 1999; Tarchini et al. 2000; Song et al. 2002), as well as at the *Xmwg644* and the *Rph7* loci between rice and barley (Dubcovsky et al. 2001; Brunner et al. 2003). Considering the large number of events detected in 193 kb at the *Lr10* locus, we conclude that local rearrangements are one of the major driving force for genome evolution and contibute significantly to the mosaic pattern of conservation between grass genomes.

A duplication in the rice genome revealed by comparative in silico mapping with wheat

The results presented here suggest that a large and segmental duplication between the rice chromosomes 1L and 5S has occurred during the evolution of the rice genome. Similar previous

observations at the genetic and physical level have indicated the presence of large duplications involving the long arms of chromosomes 1 and 5 (Kishimoto et al. 1994) and the short arms of chromosomes 11 and 12 (Wu et al. 1998). Recently, numerous duplicated genes and chromosomal duplications in rice genome have also been detected by comparison of more than 2,000 mapped rice cDNA (Goff et al. 2002). Thus, the presence of several large duplications indicates that the rice chromosomes 1 and 5 have a common ancestor and a complex evolutionary history. Subsequent to a putative chromosome duplication of the ancestor of the rice chromosomes 1 and 5, extensive rearrangements may explain the distribution of duplicated blocks between the two chromosomes. In the future, the complete BAC contig assembly of the rice chromosome 5 will allow a better comparison with the complete sequence of the rice chromosome 1 (Sasaki et al. 2002).

A detailed sequence analysis of the duplicated regions on rice chromosomes 1 and 5 suggest that a low number of genes were retained since the duplication of the two segments. In total, 22 sequences corresponding to annotated genes were conserved in 870 kb on chromosome 1 and 665 kb on chromosome 5. This represents a conservation of one putative gene per 40 kb on chromosome 1 and one putative gene per 28 kb on chromosome 5. Similarly to the large segmental duplications that have shaped the Arabidopsis genome (Blanc et al. 2000), the distances between conserved genes vary greatly in size, indicating the occurrence of numerous rearrangements since the duplication of the region. The considerable size variation of the duplicated segment suggests differential expansions through the activity of mobile elements. Evidence for expansion by multiple insertion of retrotransposons was found between the *ACT* and *CCF* on rice chromosomes 1 (56 kb) and 5 (6.3 kb). This represents an 8.8 fold expansion of the distance between the genes on chromosome 1 compared to the paralogs on chromosome 5.

The identification of a chromosomal duplication in the rice genome by comparative *in silico* mapping using the consensus genetic map of wheat chromosome group 1S and the physical data of wheat chromosome 1AS raises the question whether the duplication predated the separation of the two species. From our analysis, colinearity was better between wheat chromosome 1AS and rice chromosome 5S than between wheat chromosome 1AS and rice chromosome 1L. Furthermore, BLAST analysis with wheat chromosome 1AS genes and mapped ESTs gave better E-values for homologs on rice chromosome 5 than on chromosome 1. Thus, our data suggest that the segmental duplication between the rice chromosome 1L and 5S might have predated the

divergence of wheat and rice, suggesting the presence of a paralogous segment in the wheat genome. Comparative mapping of the receptor-like kinase *Lrk* and *Tak* genes in grass species have shown the presence of a specific duplication in Triticeae from the wheat chromosome groups 1 and 3 (Feuillet and Keller 1999). In addition, wheat chromosome group 3 is found colinear with rice chromosome 1 (Gale and Devos, 1998), making wheat chromosome 3 a good candidate for further investigations on colinearity.

In conclusion, *in silico* comparative analysis have revealed the presence of a novel colinear region between wheat chromosome 1AS and rice chromosome 5S. However, the low degree of conservation between orthologous segments indicates extensive genomic rearrangements.

The mosaic structure of the colinearity between wheat and rice due to gene movements and segmental duplications in rice might limit the use of the rice genome for positional cloning in wheat.

Chapter 6

Ancestral genome duplication in rice

Romain Guyot and Beat Keller

(2004)

Genome 47:610-614

6 Ancestral genome duplication in rice

6.1 Abstract

The recent availability of the pseudochromosome sequences of rice allows for the first time the investigation of the extent of intra-genomic duplications on a large scale in this agronomically important species. Using a dot-matrix plotter as a tool to display pairwise comparisons of ordered predicted coding sequences along rice pseudochromosomes, we found that the rice genome contains extensive chromosomal duplications accounting for 53% of the available sequences. The size of duplicated blocks is considerably larger than previously reported. In the rice genome, a duplicated block size of >1 Mb appears to be the rule and not the exception. Comparative mapping has shown high genetic colinearity among chromosomes of cereals, promoting rice as a model for studying grass genomes. Further comparative genome analysis should allow the study of the conservation and evolution of these duplication events in other important cereals such as rye, barley, and wheat.

6.2 Introduction

Whole-genome duplication is considered to be one of the major mechanisms in the genome evolution of eukaryotes. Evidence for ancient whole-genome duplications was found in yeast (Wolfe and Shields 1997), as well as in vertebrates (Wolfe 2001). In plants, the complete sequencing of the small diploid genome of *Arabidopsis thaliana* has revealed the occurrence of three ancestral rounds of duplications (Blanc et al. 2000; Vision et al. 2000; Simillion et al. 2002; Blanc et al. 2003; Raes et al. 2003). The high genetic colinearity of the rice genome (*Oryza sativa*) with the larger genomes of maize, barley, and wheat (Gale and Devos 1998) has promoted rice as the model genome for studying genome evolution and to support gene isolation from grass species. Initial observations at the genetic and physical map levels have shown the presence of duplicated segments between chromosomes 1 and 5 (Kurata et al. 1994) and between chromosomes 11 and 12 (Wu et al. 1998). With the availability of the draft sequence and the comparison of 2000 mapped cDNAs, large-scale duplications were also suggested to have occurred in the past (Goff et al. 2002). More recently, a systematic sequence analysis has indicated that 15% of the rice genome is found in duplicated blocks (Vandepoele et al. 2003). The recent release by TIGR (the Institute for Genome Research) of 12 rice pseudochromosome

sequences allows for the first time a comprehensive whole-genome analysis of rice duplications for the investigation of both the extent and the pattern of duplications, as well as the evolutionary history at the origin of the intra-genomic redundancies. Here, we report extensive duplication of the rice genome. Many of the identified duplications have a size larger than 1 Mb and cover more than half of the genome, suggesting an ancestral round of genome duplication in rice.

6.3 Material and methods

Rice dataset

Pseudo molecules and predicted coding sequences (CDS) were downloaded for each rice chromosome from the Institute for Genome Research Web site in October 2003 (http://www.tigr.org/tdb/e2k1/osa1/pseudomolecules/info.shtml). Approximately 56 000 genes have been annotated from 358 Mb of non-overlapping rice genome sequences using an automated rice annotation pipeline (Yuan et al. 2003). Pseudo-molecules were assembled using finished and unfinished ordered sequences of BAC and PAC clones. Because of the actual unfinished status of the rice genome sequence, physical gaps were introduced to the pseudo molecules.

Detection of duplicated blocks

Pairwise comparisons of ordered CDS from rice chromosomes were performed using the BLASTN algorithm (Altschul et al. 1997) locally installed, with a cut-off for expect (*E*) values of 10^{-10}. Results were parsed using the Genome Pixelizer parser (with default values) and displayed using a dot-matrix plotter: the GenoPix2D software (Cannon et al. 2003). Duplicated blocks were easily identified by visually checking for large-scale diagonals on the graphical output given by the GenoPix2D software. In some chromosomal alignments, background produced by numerous predicted CDS coding for redundant proteins belonging to repeated elements (transposases, polyproteins, etc.) were removed to refine the resolution of duplicated blocks. Fine scale analyses of duplications were performed by dot-plot alignment (Sonnhammer and Durbin 1995). Supplemental data are available at our Web site (http://www.unizh.ch/botinst/molec_website/rice).

6.4 Results and discussion

The recent detailed analysis of a small duplicated region in rice, the distal part of rice chromosomes 1L and 5S, suggested that conservation is strictly limited to coding regions (Guyot et al. 2004). This was also observed in the duplicated regions of the *Arabidopsis* genome (Blanc et al. 2000). Thus, we assumed that conservation in duplicated blocks might be limited to coding sequences in the rice genome and we focused our analysis on ordered, predicted coding sequences (CDS). Fifty-six thousand predicted and ordered CDS from the 12 rice pseudochromosomes released by TIGR were analysed by a map-based approach to identify duplicated chromosomal blocks. The ordered CDS from each pseudochromosome were compared with each other using the BLASTN algorithm and results were displayed with a dot-matrix plotter (Cannon et al. 2003). In total, 52 duplicated blocks located on all 12 pseudochromosomes were identified, accounting for 189.95 Mb and covering 52.9% of the rice genome sequence (Fig. 1). This level of duplication is significantly higher than the 15% of duplicated blocks recently found using a dataset of 1025 genomic contigs assembled from ~70% of the annotated rice genome (Vandepoele et al. 2003). Close analysis of duplications revealed that the block sizes range from 0.12 Mb (block 2-6-A) to 16.76 Mb (block 1-5-C) and that blocks with considerable size are common: 30 blocks (57%) have a size of >1 Mb and 8 macroblocks (15%) have a size of >9 Mb (Table 1). The block sizes in rice are significantly larger than in *Arabidopis*, where the majority of blocks are <1 Mb and where the largest blocks have a size of ~2 Mb (Simillion et al. 2002).

As recently observed for a duplication between rice chromosomes 1 and 5 (blocks 5-1/1-5-A, Table 1) (Guyot et al. 2004), there are considerable size differences among rice duplicated block pairs, ranging from a size difference of 0.05 Mb (blocks 2-6-A/6-2-A) to 7.79 Mb (blocks 5-1-C/1-5-C). A pairwise comparison between chromosomes 2 and 6 has revealed the presence of three duplicated blocks of CDS between these two chromosomes (blocks 2-6/6-2-A, -B, and -C), representing 12.12 and 20.29 Mb of genomic sequences, respectively. The comparison between blocks 2-6/6-2-B and 2-6/6-2-C indicated a size difference of 3.24 and 4.68 Mb, respectively (Table 1). These great size differences among duplicated segments suggest the occurrence of rearrangements and differential expansion through the activity of mobile elements since duplication of the region. A further detailed analysis is necessary to elucidate precisely the origin

of this large size disparity between block pairs in the rice genome. Additionally, the analysis of duplications involving rice chromosome 2 has revealed the presence of a supplementary duplicated block of 10.92 Mb that is conserved in chromosome 4. The presence of this extra block illustrates a patchwork distribution of duplicated blocks between the different rice chromosomes. A remarkable conservation was also observed between chromosomes 10 and 3. Seventy-four percent of pseudochromosome 10 (six blocks altogether accounting for 16.6 Mb) are conserved as duplicated blocks in the short arm of chromosome 3 (Fig. 1). Two others blocks on chromosome 3 (11.38 Mb) are also duplicated with chromosome 7. Again, this patchwork distribution of duplicated blocks suggests numerous and complex intra- and interchromosome rearrangements, as well as block inversions and translocations. Moreover, an ancestral event of interchromosome fusion or breakage is suggested by the presence of a significant part of chromosome 10 within the short arm of chromosome 3. Our studies provide evidence for a massive genome reorganization and a divergent evolution of block pairs after duplication indicated by (*i*) the patchwork distribution of 52 blocks scattered along all rice chromosomes, (*ii*) some duplicated blocks found in opposite orientations with respect to the centromere, and (*iii*) considerable size differences among these block pairs. However, despite these numerous rearrangements, macro blocks were maintained that indicate a different situation when compared with *Arabidopsis*, in which an extensive reshuffling involving small blocks has completely remodelled the genome structure (Blanc et al. 2000). We conclude that the structure of the *Arabidopsis* and rice genomes have been differentially modified after a whole-genome duplication. Thus, the limited colinearity found so far between the plant model genomes of *Arabidopsis* and rice reflects a differential and complex paleo-polyploid history of these two model genomes. A surprising observation was the overlapping location of different duplicated blocks (Fig. 1, red arrowheads). In the *Arabidopsis* genome, the presence of overlapping duplicated blocks indicated multiple rounds of large-scale duplications (Vision et al. 2000; Blanc et al. 2003). Although additional detailed analyses are required, the presence of such blocks indicates that the evolutionary history of the rice genome is likely more complex than expected. Dating attempts have estimated the age of a putative rice genome duplication from between ~40 and 50 million years ago (Goff et al. 2002) to ~70 million years ago (Vandepoele et al. 2003). In both cases, duplications predate the divergence between *Bambusoideae* (*O. sativa*) and *Pooideae*, indicating that chromosomal duplications observed in rice might be conserved in the genomes of

agronomically important species, such as barley, rye, and wheat. A better understanding of the genomic history of rice, the model organism for grasses, will give new insights into the mechanisms of genome evolution and should allow to study the conservation and evolution of duplication events in other important cereals such as rye, barley and wheat.

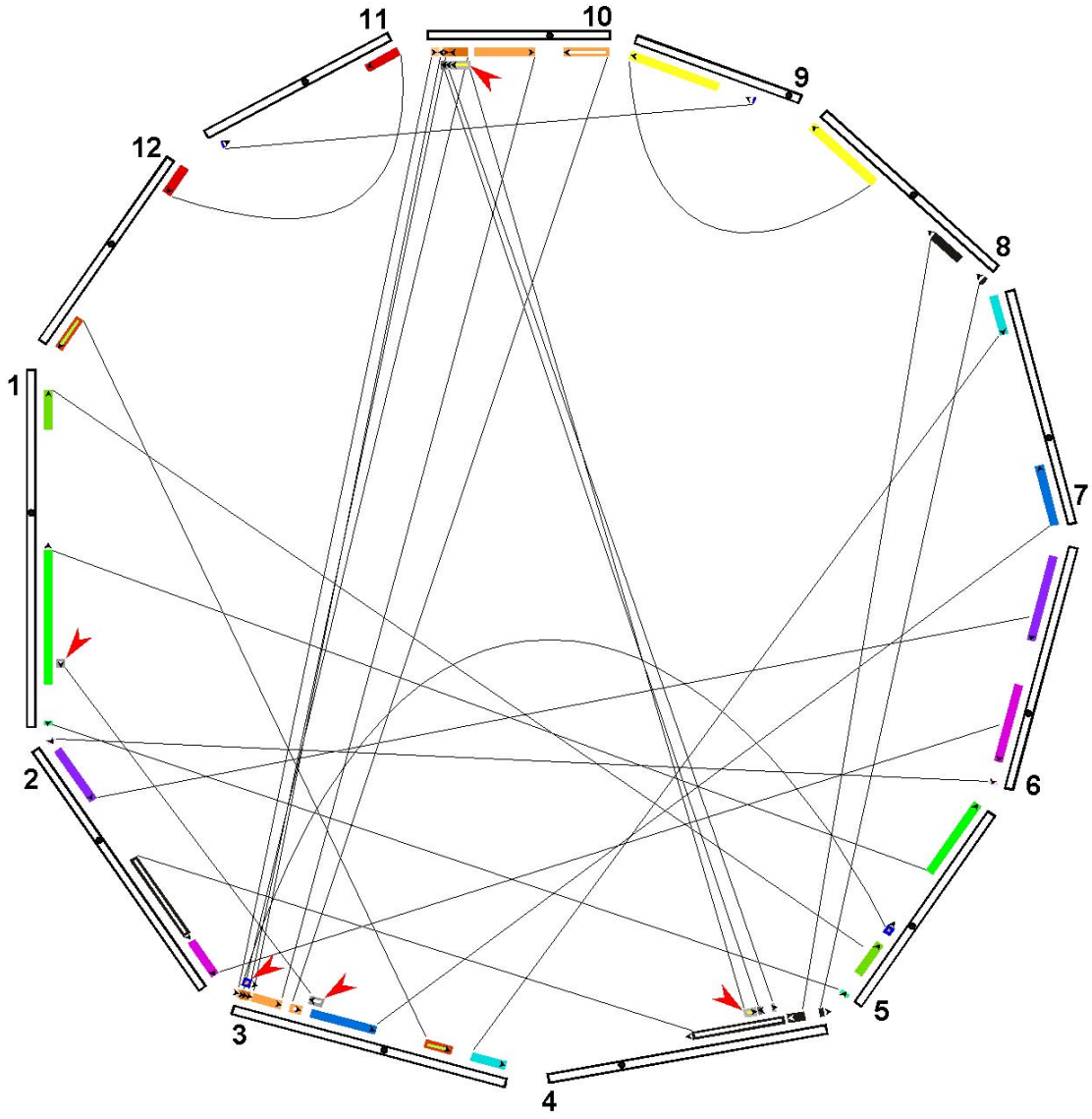

Figure 1 Genomic map of duplicated blocks in the rice genome
Coloured blocks indicate the position of conserved regions in different rice chromosomes. BAC/PAC clones flanking the duplicated regions are reported in Table 1. The lines link duplicated blocks. Black arrowheads indicate the relative orientation of duplicated blocks and red arrowheads indicate regions where overlap between duplicated blocks was observed.

Table 1 Identified duplicated blocks in the rice genome

Duplicated blocks are named according to their chromosomal location (number #1), to the chromosomal location of the duplicated block (number #2) and to the number of duplicated blocks found within the chromosome (letter #A). CDS flanking the duplicated blocks are indicated using the TIGR nomenclature (position 1 & position 2) and the accession number of BAC/PAC clones carrying such CDS is also given. Positions of flanking CDS and the size of duplicated blocks is reported according to the sequence of the rice chromosome pseudomolecules from TIGR. The size difference (Mb) between duplicated blocks is also reported.

Block	Chromosome	Flanking CDS on block	CDS position 1 (Mb)	CDS position 2 (Mb)	BAC accession of CDS position 1	BAC accession of CDS position 2	Block sizes (Mb)	Duplicated block	Size difference
1-3-A	1	5735-5897	35.15	36.12	AP003734	AP003251	0.97	3-1-A	0.54
1-5-A	1	6866-6952	42.05	42.49	AP003263	AP003277	0.44	5-1-A	0.26
1-5-B	1	553-1233	3.05	7.39	AP002539	AP002481	4.34	5-1-B	0.12
1-5-C	1	6271-3552	38.41	21.65	AP003431	AP003334	16.76	5-1-C	7.79
2-4-A	2	2832-4508	17.65	28.57	AP005285	AP004129	10.92	4-2-A	0.59
2-6-A	2	36-56	0.17	0.29	AP004187	AP004851	0.12	6-2-A	0.05
2-6-B	2	305-1434	1.73	8.93	AP004997	AP006160	7.2	6-2-B	3.24
2-6-C	2	5389-4586	33.93	29.13	AP004240	AP003983	4.8	6-2-C	4.68
3-10-A	3	74-15	0.46	0.08	AC113930	AC125411	0.38	10-3-A	0.51
3-10-B	3	262-343	1.6	2.05	AC144426	AC118132	0.45	10-3-B	0.18
3-10-C	3	253-176	1.54	1.06	AC144426	AC118980	0.48	10-3-C	0.12
3-10-D	3	86-173	0.51	1.01	AC105733	AC118980	0.5	10-3-D	1.88
3-10-E	3	936-351	5.58	2.11	AC132214	AC118132	3.47	10-3-E	3.63
3-10-F	3	1126-1321	6.88	8	AC135157	AC134240	1.12	10-3-F	4.47
3-12-A	3	3650-4336	21.97	26.55	AC109601	AC105747	4.58	12-3-A	1.19
3-1-A	3	1700-1458	10.38	8.87	AC137931	AC139168	1.51	1-3-A	0.54
3-5-A	3	1515-1644	0.55	1.49	AC079356	AC130608	0.94	5-3-A	0.18
3-7-A	3	4868-5576	29.68	33.57	AC135225	AC133339	3.89	7-3-A	3.67
3-7-B	3	1598-2803	9.72	17.21	AC084405	AC135598	7.49	7-3-B	2.53
4-10-A	4	4535-4596	28.11	28.47	OSJN00072	OSJN00188	0.36	10-4-A	0.06
4-10-B	4	4300-4254	26.52	26.24	OSJN00201	OSJN00102	0.28	10-4-B	0.11
4-10-C	4	4002-4234	24.72	26.1	OSJN00273	OSJN00112	1.38	10-4-C	0.84
4-2-A	4	2912-4755	17.9	29.41	OSJN00267	OSJN00077	11.51	2-4-A	0.59
4-8-A	4	5501-5577	33.93	34.4	OSJN00092	OSJN00099	0.47	8-4-A	0.31
4-8-B	4	5128-4757	31.7	29.42	OSJN00250	OSJN00077	2.28	8-4-B	1.93
5-1-A	5	55-83	0.26	0.44	AC104285	AC129716	0.18	1-5-A	0.26
5-1-B	5	575-1203	3.52	7.74	AC124144	AC118286	4.22	1-5-B	0.12
5-1-C	5	3033-4615	19.23	28.2	AC137617	AC130728	8.97	1-5-C	7.79
5-3-A	5	105-270	9.23	9.99	AC137698	AC137075	0.76	3-5-A	0.18
6-2-A	6	80-46	0.44	0.27	AP001129	AP001129	0.17	2-6-A	0.05
6-2-B	6	4629-2997	28.78	18.34	AP003621	AP004730	10.44	2-6-B	3.24
6-2-C	6	589-2090	3.22	12.7	AP002536	AP004811	9.48	2-6-C	4.68
7-3-A	7	83-1338	0.47	8.03	AP003759	AP003812	7.56	3-7-A	3.67
7-3-B	7	4740-3902	29.2	24.24	AP005199	AP004230	4.96	3-7-B	2.53
8-4-A	8	16-40	0.1	0.26	AP005406	AP003909	0.16	4-8-A	0.31
8-4-B	8	656-1379	4.15	8.36	AP005606	AP004657	4.21	4-8-B	1.93
8-9-A	8	2938-4386	18.3	27.46	AP003947	AP004623	9.16	9-8-A	1.54
9-11-A	9	902-883	5.58	5.45	AP006464	AP005907	0.13	11-9-A	0.07
9-8-A	9	1634-3355	10.39	21.09	AP005508	AP006459	10.7	8-9-A	1.54
10-3-A	10	3225-3373	20.55	21.44	AE017117	AE017120	0.89	3-10-A	0.51

Chapter 7

Gene inactivation and gene loss are the major driving force in the evolution of gene-dense *β-galactosidase* loci in Triticeae.

Romain Guyot[‡], Bernadette Von Malek[‡], Edith Schlagenhauf and Beat Keller[†]

[‡] These authors have contributed equally to the work.

Manuscript in preparation

7 Gene inactivation and gene loss are the major driving force in the evolution of gene-dense *β-galactosidase* loci in Triticeae

7.1 Abstract

To study genome evolution in grasses, we analyzed the gene content of two paralogous and one orthologous segment at the barley and diploid wheat (*T. monococcum*) *β-galactosidase (β-GAL*) loci. We found an organization of genes in islands, with a high gene density which is similar to that of Arabidopsis. Comparison of two paralogous segments in barley indicated an ancestral origin through a segmental duplication followed by a movement of the duplicated segment from barley chromosome 5HL to chromosome 7HS. Comparative sequence analysis revealed mechanisms of paralogous gene inactivation such as deletion, mutation and transposon insertion in coding regions. Comparisons between barley paralogous regions and the rice genome revealed partial micro-colinearity interrupted by non-conserved coding regions, giving a mosaic organization of conserved sequences. In addition, the translocation of the paralogous segment results in a perturbation of colinearity between barley chromosome 7HS and the rice genome. Our data indicate that gene inactivation and gene loss are major factors in rapid evolutionary divergence of orthologous loci in Triticeae genomes.

7.2 Introduction

Grasses evolved from a common ancestor living about 77 MYA (Kellogg 2001; Gaut 2002) and later separated into the *Ehrhartoideaee* (e.g. rice), the *Pooideae* (including wheat and barley) and the *Panicoideae* (e.g. maize and sorghum) sub-families. The divergence between rice and the *Pooideae* is estimated at 50 MYA. During the last 25 million years, the diversification of the *Pooideae* sub-family was at the origin of the Triticeae tribe. In Triticeae, diploid wheat and barley species (*Hordeum vulgare*) diverged 13 MYA (Gaut 2002), whereas bread wheat (*Triticum aestivum*) was the result of successive and more recent hybridizations. Bread wheat is an allohexaploid species carrying three different homoeologous genomes (A, B and D), that are related to a common diploid ancestor living 2.5-4.5 million years ago (Huang et al. 2002).

Comparative genomic analysis is providing the basis to understand the evolution of grass genome organization. Initial comparisons based on genetic mapping revealed that homoeologous wheat chromosomes as well as chromosomes of closely related species such as barley and *T.*

monococcum (diploid wheat, A^m genome) are remarkably conserved at the macro-level (Devos et al. 1993; Dubcovsky et al. 1996) with few chromosomal rearrangements (Nelson et al., 1995). Despite large variation in genome size and organization in the grass family, comparative studies have also demonstrated a significant conservation of markers and gene order along chromosomes of distantly related grass species such as maize, rice and wheat (Ahn et al. 1993; Kurata et al. 1994), and between rice and Triticeae species (VanDeynze et al. 1995a; VanDeynze et al. 1995b). A consensus grass map aligning the genomes of twelve different species was drawn using rice as a reference, which is the smallest genome among grasses (Gale and Devos 1998). Conservation of the macro-colinearity has promoted rice as a model species to study genome evolution in grasses and as a reference genome for positional cloning in species with larger genomes (Peng et al. 1999; Keller and Feuillet 2000). However, the conservation of the gene order and orientation at the mega-base level remains critical for efficient utilization of a model species for the development of markers and for the identification of regions that might contain candidate genes for the trait of interest. Following the progress in the isolation and the sequencing of large stretches of genomic DNA in Triticeae, the direct comparison of orthologous sequences between rice and Triticeae has shown that the disruption of the micro-colinearity is the rule and not the exception. In addition to an intense accumulation of transposable elements in Triticeae genomes, numerous non-conserved genes between orthologous loci generate a complex mosaic conservation of the colinearity (Feuillet and Keller 1999; Li and Gill 2002; Brunner et al. 2003; Yan et al. 2003; Guyot et al. 2004). Gene amplification, gene movement and gene loss mechanisms were suspected to account for the majority of the non-collinear sequences present in orthologous regions (Song et al., 2002). This frequent disruption of the micro-colinearity may complicate the use of rice as a model for map-based cloning and suggest a dynamic genome evolution in Triticeae.

Here, we report molecular details of the duplication and the translocation of the *β-GAL* locus in barley at the basis of local perturbations of colinearity between the barley, wheat and rice genomes. Comparative analysis revealed that paralogous gene inactivation is one of the major evolutionary forces at the origin of the mosaic conservation of genes at the *β-GAL* loci.

7.3 Materials and methods

Genetic mapping in barley

Genetic mapping of the ß-gal loci in barley was performed on a small segregating F_2 population derived from a cross between Cebada Capa and Bowman (Brunner et al. 2003). DNA isolation, southern blotting and hybridisation were performed as described previoulsy (Graner et al. 1990). Linkage analysis was based on sixty-five polymorphic molecular markers derived from barley (BCD, MWG, ABC and ABG), wheat (PSR) and oat (CDO; (Neu et al. 2003)). Mapping was performed using Mapmaker, version 3.0b (Lander et al. 1987).

BAC clone isolation and sequencing

Screenings of the *T. monococcum* cv DV92 (Lijavetzky et al. 1999) and the *Hordeum vulgare* cv Morex BAC libraries (Yu et al. 2000), BAC DNA preparation for fingerprint analysis and BAC end sequencing were performed as described previously (Stein et al. 2000). Preparation for shotgun cloning and "low-pass" sequencing of the barley BAC clones 343I11 (448 sequences) and 124H14 (336 sequences) was also carried out as described previously (Stein et al. 2000). Assembly of shotgun sequences were carried out as described earlier (Wicker et al. 2003b). Thirty three contigs and thirteen singletons, accounting for 49.4 kb of DNA sequence were generated for 343I11. Eleven contigs, accounting for a total length of about 33 kb of sequences were generated for 124H14. Sequences of the assembled contigs were deposited in EMBL under the following accession numbers: 343I11-1, AJ717740; 124H14-1, AJ717741; 65B22-1, AJ717742

Southern hybridization

Hybridization experiments were performed on 124H14 and 65B22 barley BAC clone DNA, digested with the restriction enzyme *Hind*III. Probes were designed as inserts of five different shotgun clones of 343I11 showing similarities to *Upr-343I11*, *Kin-343I11*, *Skp1-343I11*, DHQ_*synthase-343I11* and *Aip-343I11* genes (Table 1). Hybridization was performed as described previously (Brunner et al. 2003).

Sequence analysis

DNA sequences were analyzed using BLAST algorithms (Altschul et al. 1997) against public and local sequence databases. Detailed analysis was performed with the GCG Sequence Analysis Software Package version 10.1 (Madison, WI) and by dot plot (Sonnhammer and Durbin 1995). Coding regions were determined by a combination of gene prediction software (RiceGAAS, http://ricegaas.dna.affrc.go.jp) and amino-acid and nucleotide alignments. Intron and exon structure of predicted genes were confirmed using alignment against rice full-length cDNA sequences (Kikuchi et al. 2003b). Sequence alignments were performed using GAP (GCG) with a gap creation penalty of one and a gap extension penalty of zero. tRNA genes were identified using the tRNAscan-SE engine (Lowe and Eddy 1997). Putative repetitive elements were first annotated by amino-acid and nucleotide similarity searches against repeat databases: Repbase (Jurka 2000) and TREP: the Triticeae Repeat sequence databases (Wicker et al. 2002) using an E-value of $E < 10^{-10}$. The ends of repetitive elements were carefully investigated using dot plot pairwise alignments of the query sequence against itself or against genomic sequences available in public databases.

Data and map sources

The consensus map of barley was downloaded from the GrainGenes web site (http://wheat.pw.usda.gov/ggpages/maps.shtml). The sequences of RFLP probes located on the barley genetic map and used for comparative mapping among grasses were downloaded from the NCBI (http://www.ncbi.nlm.nih.gov/) and from GrainGenes (http://wheat.pw.usda.gov/index.shtml) web sites. The nucleotide sequences of retrieved RFLP probes were used for BLAST searches (Altschul et al. 1997) with a threshold for the *E* value of $E < 10^{-20}$. The sequences of *O. sativa* ssp. *japonica* BAC clones were downloaded from the RiceGAAS web site (ftp://ftp.dna.affrc.go.jp/pub/RiceGAAS), as of April 2004. Triticeae EST sequences assigned by chromosome bin mapping were downloaded from the GrainGenes web site http://wheat.pw.usda.gov/NSF/progress_mapping.html.

7.4 Results

Sequence analysis of *β-GAL* gene regions in barley and diploid wheat

Initial analysis of shotgun sequences from the two barley BACs 124H14 and 343I11 revealed sequences similar to plant *β*-galactosidase genes. Sub-clones containing *β-GAL* genes were then isolated and sequenced. Complete sequence assembly of sub-clones and the assembly of BAC shotgun sequences into contigs was sufficient to reveal the genes present on the BACs. In barley BAC 124H14 six coding regions were identified in four different unordered contigs and in barley BAC 343I11 seven coding regions were identified in seven different unordered contigs (Table 1). These barley coding regions were classified into two categories. The *β-GAL, tRNA* and *Upr* genes were conserved in both barley BACs. The other identified coding regions were found either specific to 124H14 (*ISP4-124H14*, *FGT-124H14, Sin-124H14*) or to 343I11 (*Skp1-343I1, Kin-343I11, Aip-343I11*) (Table 1 and Figure 1 A).

The sequencing of a sub-clone surrounding a *β-GAL* gene in the *T. monococcum* BAC 65B22 revealed the presence of three genes over a distance of 10.5 kb. Two of them (*β-GAL-65B22* and *tRNA-65B22*) were found similar to genes identified in barley BACs 124H14 and 343I11 (Table 1 and Figure 1 A). Southern hybridization experiments were carried out on BAC 65B22 using coding regions of BAC 343I11 as probes. The three coding sequences *Upr-343I11*, *Skp1-343I11* and *Aip-343I11* gave a signal on DNA of BAC 65B22 (data not shown). Thus, six, seven and six coding regions were respectively identified over an estimated BAC size of 40 kb for 124H14, 80 kb for 343I11 and 50 kb for 62B32, revealing a high gene-density. In the completely sequenced contigs 343I11-1, 124H14-1 and 65B22-1 (Figure 1A, Table 1), genes are organized in islands in which the gene density ranges from 1.8 to 3.5 kb/gene. BLASTN searches using all contigs as queries revealed the presence of several transposable elements. In 124H14, the CACTA transposon *Caspar* (Wicker et al. 2003a) was found inserted within the third intron of the *Upr-124H14* gene of the 124H14-1 contig (Figure 1 A). Additional fragments of *Caspar* as well as sequences similar to LTR (Long Terminal Repeat) of the retrotransposon *BARE-1* (Vicient et al. 1999) were identified in the remaining contigs and singletons (data not shown). Transposable elements belonging to the class I retrotransposon (*BARE-1, Usier, Sabrina, Sukkula and Yvonne*), class II MITES (*Thalos* and *Hades*) and to an unclassified element were identified in 343I11, but no sequences were similar to *Caspar* (data not shown).

Identification of micro-colinearity between *T. monococcum* and barley sequences

At the BAC level, five coding regions (*β-GAL*, *tRNA*, *Upr*, *Skp1* and *Aip*) were found conserved between *T. monococcum* and the barley BAC 343I11 (Figure 1 A). Only three genes (*β-GAL*, *tRNA* and *Upr*) are conserved between *T. monococcum* and barley 124H14 and between the two barley BACs (Figure 1 A). This result suggests that the conservation between *T. monococcum* and the barley BAC 343I11 is better than between the other fragments. A detailed comparative sequence analysis was carried out using the largest and gene-dense contigs established in barley and *T. monococcum* BACs (contigs 343I11-1, 124H14-1 and 65B22-1 respectively, Table 1).

Table 1 Predicted coding regions in the BAC clones 65B22 from *T. monococcum* and 343I11 and 124H14 from barley

Locus	BAC	Contig #	Contig size (bp)	Gene name	Predicted function	BLASTX
Tm- β-GAL	65B22	65B22-1	10571	***Glyco_hydro-65B22***	Putative glycosyl transferase (p)	AAN65437 (*O. sativa*) E=10^{-49}
				β-GAL-65B22	Putative glycosyl hydrolase family 35	BAD08952 (*O. sativa*) E=0.0
				tRNA-65B22	tRNA type : Pro, Anticodon : TGG	ND
Hv- β-GAL1	124H14	124H14-1	10640	***β-GAL-124H14***	Putative glycosyl hydrolase family 35	BAD08952 (*O. sativa*) E=0.0
				tRNA-124H14	tRNA *	ND
				Upr-124H14	Unknown protein (p) *	NP_179694 (*A. thaliana*) E=10^{-16}
				TNP2-like-124H14	TNP2-like transposase protein (p)	AAL73531 (S. *bicolor)* E=10^{-44}
		124H14-2	1332	***Isp4-124H14***	Similar to Isp4-like rice protein (p)	NP_910351 (*O. sativa*) E=10^{-28}
		124H14-3	5958	***Isp4-124H14***	Similar to Isp4-like rice protein (p)	NP_910351 (*O. sativa*) E=0.0
				FGT-124H14	Similar to flavonol 3-O-glucosyltransferase (p)	NP_915669 (*O. sativa*) E=10^{-20}
		124H14-4	2371	***Sin-124H14***	Similar to Put. salt-inducible protein (p)	NP_910350 (*O.sativa*) E=10^{-48}
				FGT-124H14	Similar Put. flavonol glucosyltransferase (p)	NP_916450 (*O. sativa*) E=10^{-16}
Hv- β-GAL2	343I11	343I11-1	7349	***DHQ_synthase-343I11***	Dehydroquinate synthase (p)	AAL77575 (*L. esculentum*) E=10^{-29}
				β-GAL-343I11	Glycosyl hydrolase family 35 *	BAD08952 (*O. sativa*) E= 10^{-100}
				tRNA-343I11	tRNA, type : Pro, Anticodon : TGG	ND
				Upr-343I11	Unknown protein (p)	NP_179694 (*A. thaliana*) E=10^{-35}
		343I11-2	1286	***DHQ_synthase-343I11***	Dehydroquinate synthase (p)	AAM61355 (*A. thaliana*) E= 10^{-27}
		343I11-3	1372	***Upr-343I11***	Unknown protein (p)	ND
		343I11-4	1112	***Skp1-343I11***	Kinetochore protein (P)	CAE53885 (*T. aestivum*) E=10^{-24}
		343I11-5	482	***Skp1-343I11***	Kinetochore protein (P)	AAD34458 (*M. sativa*) E=10^{-15}
		343I11-6	1691	***Kin-343I11***	Receptor-protein kinase (P)	CAE03769 (*O. sativa*) E=10^{-117}
		343I11-7	1911	***Aip-343I11***	Auxin Induced (aldo/keto red.) protein (P)	P49249 (*Z. mais*) E=10^{-20}

ND : Not determined, (p) partial gene, * pseudogene.

The three contigs share a minimum of two coding regions conserved in the same order and orientation (*β-GAL* and *tRNA* genes) and three coding regions (*β-GAL, tRNA* and *Upr*) were found colinear between the two barley fragments (Figure 1 A).

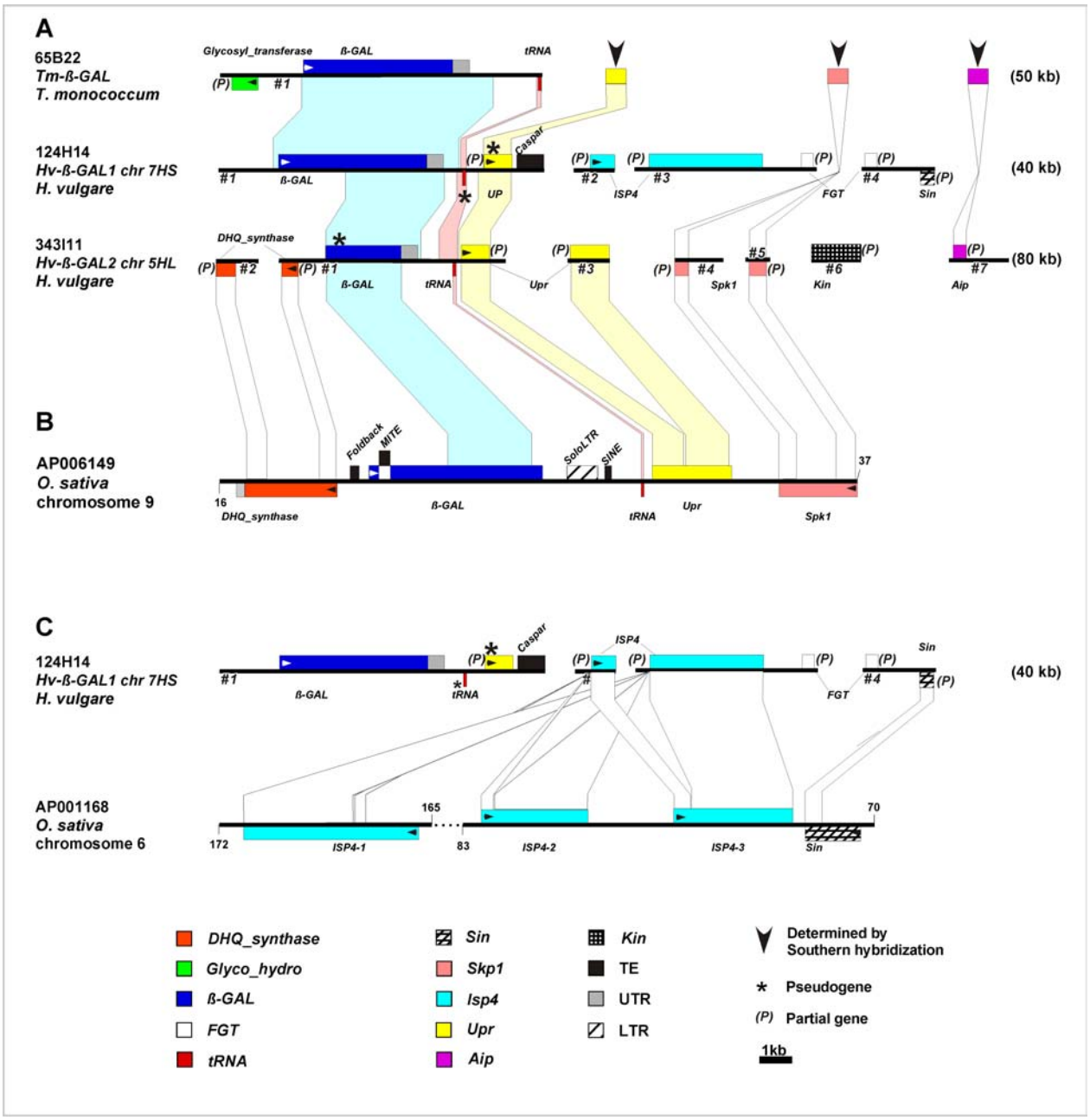

Figure 1 Schematic representations of the micro-colinearity between barley and *T. monococcum* BACs and the rice genome

A. Schematic representation of the gene content of the Triticeae BACs and the micro-colinearity between the 65B22 (*T. monococcum),* 124H14 and 343I11 (*H. vulgare*) BACs. Coding regions of genes and transposable elements are represented by colored boxes. Grey boxes represent UTR regions defined by similarities searches against ESTs databases. Asterisks indicate pseudo-genes and (p) indicate genes partially covered by sub-clones. Genes are named as in Table 1. Estimation of the BAC sizes is indicated on the right (kb). In the *T. monococcum* BAC 65B22, black arrowheads indicate the presence of genes determined by Southern hybridization. Distances and order of contigs are arbitrary.

B. Schematic representation of the micro-colinearity between the barley and *T. monococcum* BACs and a segment of the rice AP006149 BAC on chromosome 9L.
C. Schematic representation of the micro-colinearity between the barley BAC 124H14 and the rice BAC AP001168 on chromosome 6S.

The partial *Glyco_hydro-65B22* gene located in the proximal part of the 65B22-1 contig as well as an uncharacterized region of 1,617 bp located between *β-GAL-65B22* and *tRNA-65B22* genes (positions 8,633-10,250) were not found conserved neither in 124H14-1 nor in 343I11-1. The partial *DHQ-synthase-343I11* gene was not present neither in 124H14-1 nor in 65B22 and the non-coding part upstream the *β-GAL-124H14* gene was not conserved in 343I11-1 and in 65B22-1. These results suggest that the micro-colinearity between the three Triticeae contigs is locally interrupted by the presence of non-conserved genes. Around the *β-GAL* genes, the nucleotide conservation ranged from 82% of identity over 3.4 kb between 124H14-1 and 343I11-1, 84% of identity over 3.4 kb between 343I11-1 and 65B22-1 and up to 87% of identity over 6 kb between 124H14-1 and 65B22-1. The conserved region in barley that encompassed the *Upr* genes shows 78% of identity over a distance of 1.1 kb. Short regions of 80 bp in the promoter of the *β-GAL* genes as well as 500 bp downstream *β-GAL* between 65B22-1 and 124H14-1 and 200 bp in the promoter of the *Upr* genes between 124H14-1 and 343I11-1 were found conserved outside the coding regions. These results indicate that the conservation of the micro-colinearity is limited to coding regions and short extensions into intergenic regions. In summary, our results suggest an orthologous relationship between the barley 343I11 and the *T. monococcum* 65B22 contigs and a paralogous relationship between the two barley 343I11-1 and 124H14-1 contigs.

Pseudogenes in barley

Comparisons of orthologous and paralogous genes in the *β-GAL* region indicate differences in the gene structures. While *β-GAL-65B22* and *tRNA-65B22* genes were identified as complete genes in *T. monococcum*, one and two genes, respectively, were found disrupted in the barley BACs 343I11 and 124H14. In 343I11, the colinear genes *tRNA-343I11* and *UP-343I11* are predicted to be complete. In contrast, sequence alignments revealed a deletion of eight exons as well as stop codons in the reading frame of the *β-GAL-343I11* gene. This result indicates that *β-GAL-343I11* is a pseudogene. In 124H14-1, despite the presence of some sequence similarities with other plant tRNAs, the *tRNA-124H14* gene was not detected by tRNA prediction software (Lowe and Eddy 1997). Sequence analysis revealed that four nucleotides involved in the terminal stem structure of

tRNA-124H14 were deleted or mutated, creating a defective gene. Distal to the *pseudo-tRNA* gene, the insertion of the *Caspar* transposon as well as the identification of several stop codons within the reading frame of the *UP-124H14* gene indicated that *UP-124H14* is also an inactivated gene (Figure 1 A). Thus, the comparison between the barley paralogous segments revealed duplication of a total of three genes. The sequence analysis of these paralogous genes indicates that in both paralogous segments, one of these genes is predicted to be a pseudogene.

Identification of micro-colinearity between diploid wheat, barley and rice segments

BLASTN searches were conducted using the coding regions detected in barley and *T. monococcum* BACs as queries against the rice genomic sequences (*O. sativa* ssp *japonica*). Five different Triticeae genes *DHQ_synthase*, *β-GAL*, *tRNA, Upr* and *Skp1* from barley 343I11 BAC showed significant similarity to one rice BAC located in the distal part of the chromosome 9 long arm (AP006149, Supplementary data). A 20 kb segment (Os_AP006149 contig) was extracted from the 171 kb of the rice BAC AP006149 (positions 17 to 37 kb) for detailed analysis. Five putative active genes (*DHQ_synthase-oschr9*, *β-GAL-oschr9*, *tRNA-oschr9, Upr-oschr9* and *Skp1-oschr9*) were identified, annotated and used for comparative analysis.

AP006149 shares respectively five, four and three coding regions with barley BAC 343I11, with *T. monococcum* BAC 65B22 and with barley BAC 124H14 (Figure 1 B). Four Triticeae coding regions (*FGT-124H14, Kin-343I11, Aip-343I11* and *Glyco-hydro-65B22,* Table 1) were not conserved in the rice colinear region but had homologs dispersed in the rice genome (Supplementary data). These results suggest colinearity between rice and the Triticeae BACs is only partial and is disrupted by the presence of non-conserved genes.

Two coding regions identified in barley 124H14 BAC (*ISP4-124H14* and *Sin-124H14*) have homologs located in one rice BAC present on the short arm of chromosome 6 (AP001168, Supplementary data). In rice AP001168 three copies of *ISP4* genes are present (Figure 1 C). This result suggests colinearity between barley BAC 124H14 and rice BAC AP001168 located on chromosome 6. Altogether, the barley 124H14 BAC showed partial colinearity with the rice chromosomes 9 and 6.

Pairwise sequence comparison was conducted between rice BAC AP006149 (chromosome 9) and the Triticeae contigs 124H14-1, 343I11-1 and 65B22-1. The most extensive conservation was found between Os_AP006149 and 343I11-1, in which four coding regions were strictly

conserved in the same order and transcriptional orientation. 65B22-1 and 124H14-1 shared, respectively, three and two conserved regions with the rice Os_AP006149 contig, the conserved regions corresponding to coding regions. Neither introns of predicted genes nor intergenic regions and repetitive elements were found conserved between Triticeae and rice contigs, with the notable exception of the upstream regions of the *DHQ_synthase* genes. Here, a stretch of respectively 287 bp and 307 bp for 343I11-1 and Os_AP006149 were found conserved immediately upstream the start codon of both genes with 63.4 % of nucleotide identity. The intergenic distances separating colinear genes between Triticeae and rice were also compared. The comparison of distances of the *DHQ_synthase* and *β-GAL* interval showed a limited variation between barley 343I11-1 and rice Os_AP006149 contigs. A slight expansion of distance occurred in rice through the insertion of a foldback element in the promotor of the rice *β-GAL* gene. A deletion of eight exons in the *β-GAL-343I11* gene reduced distances between *DHQ_synthase* and *β-GAL* genes in barley 343I11-1 contig (Figure 1 B). Intervals between *β-GAL* and *tRNA* showed an increase in rice of approximately 2.9, 4.7 and 1.4 fold, respectively, compared to the barley contigs 124H14-1, 343I11-1 and *T. monococcum* contig (65B22). The insertion of two transposable elements in Os_AP006149 (SoloLTR type SZ-49 and SINE type F524) was partially responsible for this expansion. In contrast, distances between *tRNA* and *Upr* genes in 124H14-1 showed an increase of 2.5 and 3.2 fold respectively compared to Os_AP006149 and 343I11-1 intervals, which could not be explained by transposable element insertions. No other sequences from barley and *T. monococcum* BACs showed nucleotide identity with rice BAC AP006149.

Altogether, these data indicate a mosaic conservation of the micro-colinearity between the Triticeae BAC 343I11, 124H14 and 65B22 contigs and the rice chromosome 9 AP006149 segment and between 124H14 and the rice chromosome 6 AP001168. The micro-colinearity observed is restricted to coding sequences and no significant expansion of the distance between colinear genes was observed in Triticeae compare to rice.

Extension of the Triticeae-rice colinearity by identification of conservation between barley genetic markers, ESTs anchored in wheat bins and the rice genome

To extend the *β-GAL* loci and to identify regions of macro-colinearity between Triticeae and rice chromosomes, we genetically mapped the *β-GAL* genes from barley and *T. monococcum.* Genetic

mapping experiments revealed that 343I11 (*Hv-β-GAL2* locus) and 124H14 (*Hv-β-GAL1* locus) were located on barley chromosomes 5H and 7H, respectively. *Hv-β-GAL2* was mapped into a genetic interval (*MWG604* and *cMWG701a*) of chromosome 5HL (Figure 2A). *Hv-β-GAL1 was* mapped into a genetic interval (*MWG2018-MWG807*) of the very distal part of the barley chromosome 7H short arm (Figure 2) The *Tm-β-GAL* locus (65B22) was physically mapped on the chromosomes 5A, 5B and 5D on hexaploid wheat using Southern hybridization on a Langdon/Chinese Spring substitution lines genomic DNA (data not shown).

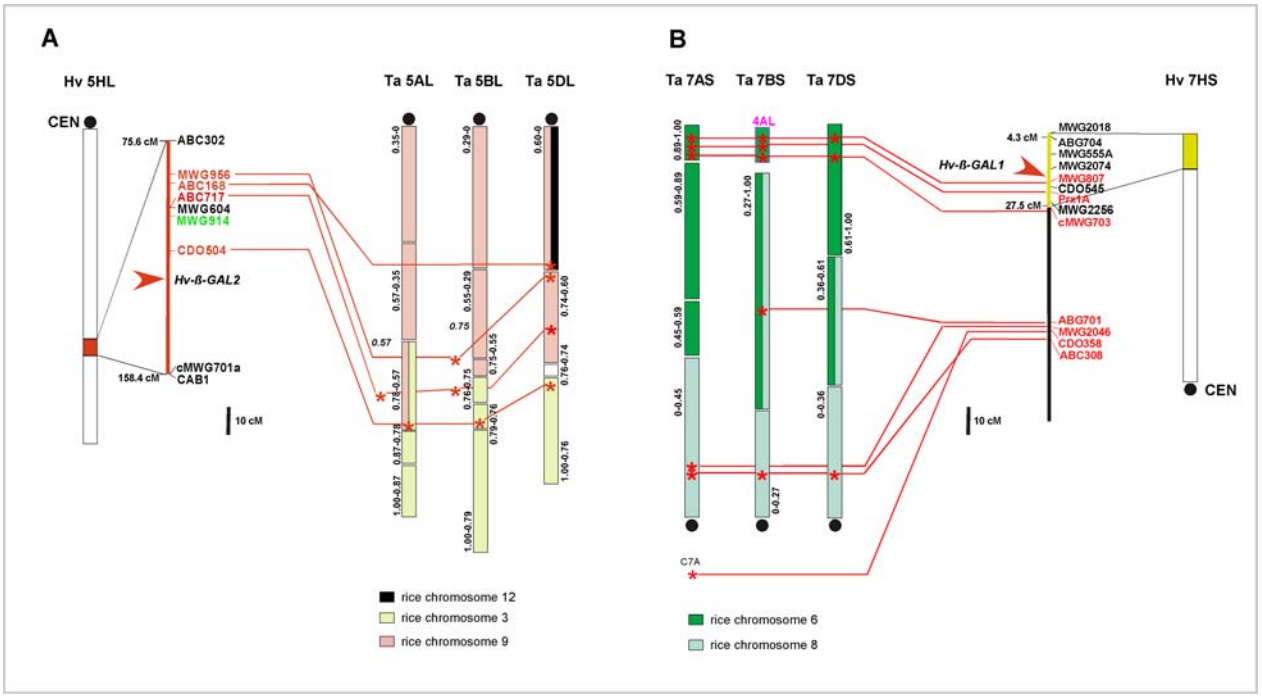

Figure 2 Comparative maps of RFLP markers from barley chromosome 5HL, barley chromosome 7HS and ESTs physically mapped in bins of hexaploid wheat chromosome groups 5L and 7S.
A. Representation of a comparative map between partial physical and genetic maps of barley chromosome 5HL and ESTs mapped in hexaploid wheat chromosomes group 5L bins. The red square on chromosome 5L indicates the physical location of a region with a high recombination rate (Kunzel 2000). The marker in green indicates sequence conservation with the rice chromosome 9 BAC AP006149.
B. Representation of a comparative map between partial physical and genetic maps of barley chromosome 7HS and ESTs mapped in hexaploid wheat chromosome group 7S bins. The yellow square on chromosome 7S indicates the physical location of a region with a high recombination rate (Kunzel 2000).
Red arrowheads indicate the genetic location of *β-GAL* genes on barley chromosomes 5HL (*β-GAL2* locus) and on barley chromosome 7HS (*β-GAL1* locus). Colored bins indicate the colinearity established between wheat chromosome bins and the rice genome by Sorrells and coworkers (Sorrells et al. 2003b; La Rota and Sorrells 2004). CEN indicate centromeres.

These results suggest that the *T. monococcum Tm- β-GAL 1 is* located on chromosome 5 and is orthologous to the barley *Hv-β-GAL2*.

The nucleotide sequences of RFLP markers surrounding the mapped *β-GAL* genes were used to study the macro-colinearity between barley chromosomes 5H and 7H and the rice genome. RFLP sequences were retrieved from public databases and used to compare with the Triticeae expressed sequence tags (EST) physically assigned into hexaploid wheat chromosome bins (Qi et al. 2003) and second with the rice genomic sequences (Guyot et al. 2004). The nucleotide sequence of 32 out of the 62 markers belonging to the *ABC302-cMWG701a* interval (75.6-158.4 cM) surrounding the location of *Hv-β-GAL2* on barley chromosome 5H were retrieved. BLASTN searches were performed against a local database composed of ESTs assigned to chromosome bins in wheat deletion lines. Six sequences showed significant identity to physically mapped ESTs. Four of them were in bins of wheat chromosome group 5L (Figure 2 A, red lines). Eleven markers (18.3%) showed significant similarities ($E<10^{-20}$) to rice genomic BAC sequences of which seven are on rice chromosome 9L. Interestingly, *MWG914* (Figure 2 A, green marker, 116 cM) was found conserved in the rice BAC AP006149 (140,719-141,190 bp), on which five coding regions were conserved with the barley BAC 343I11. Altogether these data established that the locus *Hv-β-GAL2* belongs to a large colinear region between the barley and hexaploid wheat chromosome group 5L and between the barley chromosome 5HL and the rice chromosome 9L.

The nucleotide sequences of 58 out of the 83 markers belonging to the 0-76.2 cM interval (*MWG2018-ABC308*) on barley chromosome 7H were retrieved and BLASTN searches were performed. Seven markers identified ESTs physically mapped into wheat chromosome 7S bins and three markers (*MWG807*, *Prx1A* and *cMWG703*) were also found conserved in bins of chromosome 4AL. Sixteen markers (27%) showed significant similarities ($E<10^{-20}$) to rice of which seven belong to the rice chromosome 6 short arm and five to the chromosome 8 (Figure 2 B). These data suggested that *Hv-β-GAL1* belongs to colinear regions between the short arm of the barley chromosome 7H and wheat chromosome group 7S, with the exception of the distal part of the chromosome 7BS, and between barley 7HS and a combination of rice chromosomes 6 and 8. To confirm the colinearity relationships between barley chromosomes 5HL and 7HS and rice chromosomes 9, 6 and 8, we assembled rice BACs clones surrounding the chromosome 9 BAC AP006149 and the chromosome 6 BAC AP001168.

The predicted genes were used as queries for BLASTN searches against the Triticeae ESTs mapped in wheat deletion lines. Seven overlapping rice BAC clones surrounding AP006149

(AP005092-AP005862) were assembled into a contig of 795 kb and 104 genes were predicted and extracted. Twenty-two (21%) rice predicted genes showed significant similarity to physically mapped ESTs ($E < 10^{-20}$). Thirteen of the twenty-two (59%) ESTs were assigned into bins located on long arms of chromosomes groups 5 (Figure 3) confirming orthologous relationships between the area surrounding the rice chromosome 9 BAC AP006149 and long arms of Triticeae chromosome group 5.

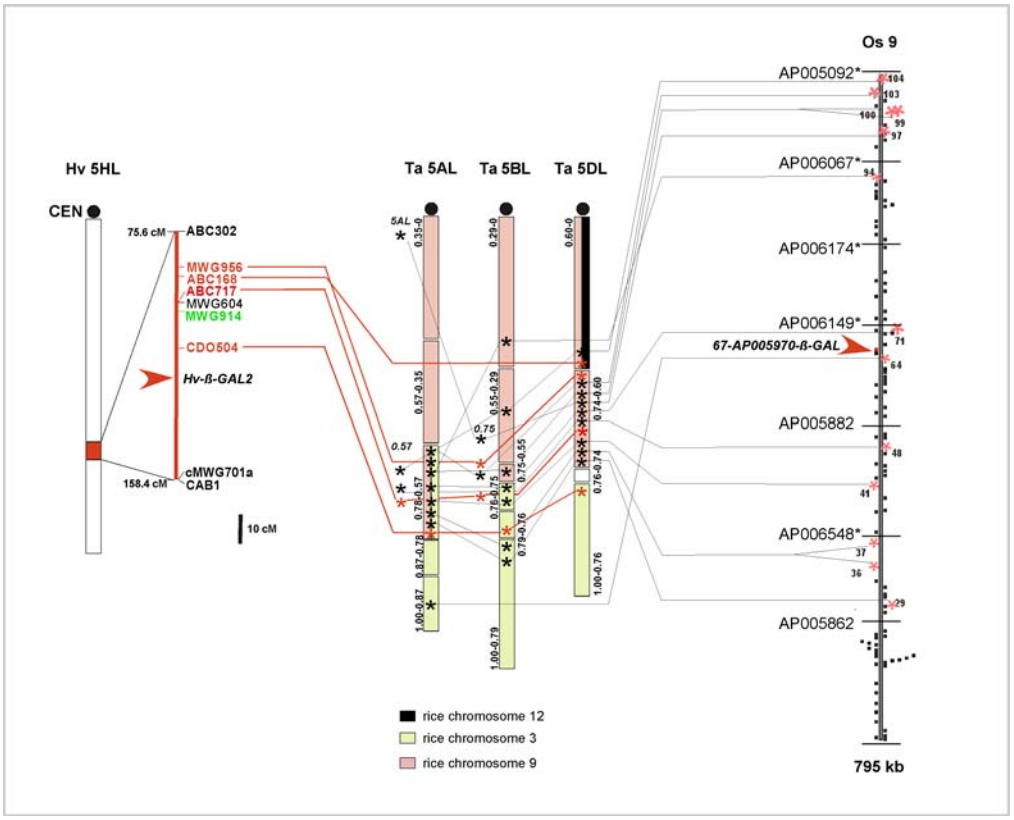

Figure 3 Comparative maps of sequences and markers conserved between the partial genetic maps of barley chromosome 5HL, ESTs mapped within bins of wheat deletion lines of chromosomes groups 5L and 104 predicted genes in a rice chromosome 9 BAC contig.
Genetic markers conserved between genetic maps and ESTs mapped in bins are indicated in red asterisks and linked with red lines.
Conservation between predicted genes in a 795 kb segment on rice chromosome 9 and ESTs physically mapped in bins of wheat chromosome group 5L are indicated by black asterisks and linked with black lines. Numbers associated with black asterisks relate to predicted genes in rice. Red arrowheads indicate the genetic location of the *β-GAL* genes on barley chromosomes 5HL (*β-GAL1* locus) and the physical location of the *β-GAL* gene on rice chromosome 9. Colored bins indicate the colinearity established between wheat chromosome bins and the rice genome by Sorrells and coworkers (Sorrells et al. 2003b; La Rota and Sorrells 2004). Black asterisks associated with accession numbers of BAC clones indicate finished sequences. CEN indicates centromere.

A similar analysis was performed in a region surrounding the rice chromosome 6 BAC AP001168. One hundred and twenty-four predicted genes were extracted from a 800 kb contig of eight BACs (AP003564-AP002542). Eight of the 124 predicted genes (6.4%) identified homologs in physically mapped Triticeae ESTs, of which only two were assigned to chromosome group 7S bins (data not shown). The six remaining ESTs homologous to rice showed a random distribution on wheat chromosomes.

7.5 Discussion

Gene organization at the *β-GAL* loci in barley and *T. monococcum*

In the Triticeae genomes, genes are not distributed randomly and there is evidence for clustering in "gene rich" regions representing probably less than 10 % of the total chromosome length (Keller and Feuillet 2000; Kunzel 2000; Sandhu and Gill 2002a).

In Triticeae, there is a large variation of the gene density ranging from about 8 kb/gene to up to 100 kb/gene (Rahman et al. 1997; Feuillet et al. 2001; Brooks et al. 2002; Chantret et al. 2004). In these sequenced regions, single genes and gene islands were found interspersed with large stretches of nested retrotransposons (Shirasu et al. 2000; Dubcovsky et al. 2001; Wicker et al. 2001; Wicker et al. 2003b). Local and differential insertions of transposable elements were shown to have an important impact on the gene organization and density.

In the present study, the two barley BACs analyzed were mapped within regions of high recombination rate located on the middle part of the chromosome 5HL and the telomeric part of 7HS (Kunzel 2000). These mapping locations and the high gene density found in the two barley BACs (124H14 and 343I11), indicate that these loci belong to gene-rich regions of the barley genome. The establishment of sequenced contigs surrounding paralogous and orthologous *β-GAL* genes in these BAC clones revealed an organization in gene-dense islands in which the density ranged from 1.8 to 3.5 kb/gene (including inactive genes and genes partially covered by contigs). The gene density observed in the three regions studied is comparable to those observed in Arabidopsis (5 kb/gene). Such high gene-density is due to an organization of genes in islands that are completely devoid of known transposable elements. However, sequence analysis revealed that transposable elements such as retrotransposons were present in the analyzed BACs outside gene islands. This observation indicates a non random suppression of insertion of transposable

elements in gene-rich regions and suggests mechanisms that may preserve gene-dense islands from transposable element insertions. Further sequencing and comparative analyses in gene-rich regions in Triticeae are required to understand the evolution of the gene-dense island structures as well as the origin of their preservation from invasions of transposable elements.

Evolution of orthologous and paralogous DNA in barley and diploid wheat

The *T. monococcum* and barley genomes are known to be closely related (Gaut 2002) and were found highly colinear (Dubcovsky et al. 1996). In our study, the barley BAC 343I11 and the *T. monococcum* 65B22 BAC were both located on chromosomes of group 5. Comparative maps between a partial genetic map of barley chromosome 5HL surrounding the location of *Hv-β-GAL2* locus (BAC 343I11) and ESTs mapped in hexaploid wheat homoelogous chromosomes group 5L bins indicated a conservation of the macro-colinearity. In addition to the general conservation of the gene content between the barley BAC 343I11 and the *T. monococcum* BAC 65B22 and the conservation of the gene order and orientation at the sequenced contig level, these data strongly indicate an orthology relationship. The conservation of the order and the orientation of three genes between the barley BACs 343I11 and 124H14 indicate a paralogy relationship. A region containing the *β-GAL* genes was likely duplicated and inserted from chromosome 5HL to 7HS. This scenario is supported by the orthology relationship established between barley 343I11 and *T. monococcum* 65B22 BACs and confirmed by a much more limited micro-colinearity between the barley 124H14 and the two orthologous 343I11 and 65B22 BACs. The presence of only one *β-GAL* gene both in *T. monococcum* and in homoeologous genomes in cv. Chinese Spring nullitetrasomic lines may indicate that this duplication/translocation event occurred in the barley lineage after the speciation between wheat and barley. However, we can not exclude the possibility that the duplication/translocation event occurred early in the diploid ancestor of Triticeae and that these events were followed by a deletion in the wheat lineage.

Our comparative analysis indicates mechanisms of paralogous gene inactivation occurring either on the ancestral or on the duplicated-derived loci at the barley paralogous *β-GAL* loci. Inactivation appears to randomly affect genes in both ancestral and duplicated-derived loci. Mechanisms of inactivation include large coding sequence deletion (*β-GAL* gene), nucleotide deletions or mutations (*tRNA* gene) and transposable element insertion (*UP* gene). Recently, we showed that CACTA transposons were highly redundant in Triticeae genomes (Wicker et al.

2003a), indicating that they play an important role in the structure and the organization of such genomes. Here, we found that *Caspar*, a CACTA transposon, disrupted the *UP* gene in the barley BAC 124H14. This result suggests that CACTA transposons may also have a mutagenic potential and they can contribute to the evolution of genes in Triticeae.

Impact of locus duplication and paralogous gene inactivation mechanisms on the evolution of grass genomes

The rice *β-GAL* locus has two strongly conserved homologous regions on barley chromosome 5HL and 7HS. Using macro-colinearity studies, there is good evidence that the *Hv-β-GAL2* locus on chromosome 5HL was ancestral and has undergone duplication and translocation on the barley chromosome 7HS. Micro-colinearity studies between the barley 124H14 BAC on chromosome 7HS and the rice genome indicate a complex colinearity with the rice chromosomes 9 and 6. We conclude that the colinearity between barley and rice is locally perturbed at the barley *Hv-β-GAL1* locus due to duplication and translocation mechanisms. Recently, comparative analysis has revealed a significant perturbation of the colinearity between wheat homoeologous chromosomes due to locus duplication, insertion and deletion (Akhunov et al. 2003). A large number of duplications and translocations in the Triticeae may have a dramatic impact on the conservation of the colinearity between Triticeae and rice genomes. Further analyses are required to understand in details the extent of such locus duplication and translocation events in the Triticeae genomes.

At the barley *β-GAL* loci, evolutionary mechanisms acting on duplicated genes are at the basis of an ongoing process of mosaic gene conservation. It is known that duplicated copies of genes are able to escape the pressure imposed by the selection allowing the accumulations of mutations in coding regions of paralogous copies. These genes are free to evolve into new functions but the most likely fates of duplicated genes are inactivation and loss (Lynch and Conery 2000). Such extensive gene loss was reported in the ancient polyploid genome of Arabidopsis (Simillion et al. 2002). Mechanisms of gene inactivation were also reported to affect homoelogous genes derived from the recent polyploidy of the wheat genome (Gu et al. 2004). Altogether, these results demonstrate that gene duplication and inactivation are major mechanisms of plant genome evolution. The main impact of differential gene loss in paralogous segments is the generation of single genes that appear to be non-colinear in orthologous regions. There are now numerous

examples of colinearity disruption between closely related grass species (Song et al. 2002; Ilic et al. 2003) and between Triticeae and rice genomes (Feuillet and Keller 1999; Li and Gill 2002; Brunner et al. 2003; Yan et al. 2003; Guyot et al. 2004). Gene movement associated with gene amplification was suspected to be at the basis of these mosaic organizations of conserved and non-conserved sequence between orthologous regions (Song et al. 2002). Our current analysis on the barley *β-GAL* loci allow us to propose that gene duplication and differential paralogous gene loss are one of the driving forces for the Triticeae genome evolution and contribute significantly to the mosaic conservation of orthologous sequences in these genomes.

Supplementary data. Detailed analysis of the BLASTN results of coding regions predicted from barley and diploid wheat BAC clones compared with rice (*O. sativa* ssp. *japonica*) genomic BAC sequences. Accession numbers of rice BAC clones, chromosome positions and BLASTN E-values are indicated.

Locus	BAC	Contig #	Gene name identified	Number of BLASTN hit with E<10^{-10}	BLASTN results
Tm-β-GAL	65B22	64B32-1	*Glyco_hydro-65B22 (p)*	2	AP005456 chr 6 $E=10^{-32}$ AC125472 chr 3 $E=10^{-32}$
			β-GAL-65B22	3	AP006149 chr 9 $E=10^{-70}$ AP003912 chr 8 $E=10^{-19}$
			tRNA-65B22	11	AB026295 chr 6 $E=10^{-28}$ AP006149 chr 9 $E=10^{-33}$
Hv-β-GAL1	124H14	124H14-1	*β-GAL-124H14*	2	AP006149 chr 9 $E=10^{-88}$ AP003912 chr 8 $E=10^{-12}$
			tRNA-124H14 *	11	AB026295 chr 6 $E=10^{-16}$ AP004788 chr 2 $E=10^{-15}$ AP006149 chr 9 $E=10^{-13}$
			Upr-124H14 (p) *	1	AP006149 chr 9 $E=10^{-13}$
			TNP2-like-124H14	ND	ND
		124H14-2	*Isp4-124H14 (p)*	1	AP001168, chr6 $E=10^{-57}$
		124H14-3	*Isp4-124H14 (p)*	2	AP001168, chr6 E=0.0 AC108761, chr9 $E=10^{-10}$
			FGT-124H14 (p)	0	ND
		124H14-4	*Sin-124H14 (p)* *FGT-124H14 (p)*	2 2	AP001168, chr6 $E=10^{-48}$ AC108499, chr5 $E=10^{-14}$
Hv-β-GAL2	343I11	343I11-1	*DHQ-synthase-343I11 (p)*	1	AP006149 chr 9 $E=10^{-45}$
			β-GAL-343I11 *	1	AP006149 chr 9 $E=10^{-13}$
			tRNA-343I11	11	AB026295 chr 6 $E=10^{-33}$ AP006149 chr 9 $E=10^{-28}$
			Upr-343I11 (p)	1	AP006149 chr 9 $E=10^{-64}$
		343I11-2	*DHQ-synthase-343I11 (p)*		AP006149, chr9 $E=5e^{-36}$
		343I11-3	*Upr-343I11 (p)*	1	AP006149, chr9 $E=10^{-11}$
		343I11-4	*Skp1-343I11 (p)*	4	AC137752, chr 11 $E=10^{-29}$ AP006149, chr9 $E=10^{-25}$
		343I11-5	*Skp1--343I11 (p)*	0	AP006149, chr9 $E=10^{-7}$
		343I11-6	*Kin-343I11(p)*	3	AC134344, chr5 $E=10^{-67}$
		343I11-7	*Aip-343I11 (p)*	1	AL606600, chr4 $E=10^{-16}$

ND : Not determined, (p) partial gene, * pseudogene.

8 General discussion

8.1 Amplification of CACTA transposons contributes to the evolution of the Triticeae genome

So far, the annotation processes of large genomic sequences were mainly focused on coding regions. Repetitive DNA that composes the main part of many cereal genomes was the subject of only few studies and is often considered as "Junk DNA" by molecular researchers. However, the accumulation of repetitive elements that can reach up to 80 % in Triticeae have an important impact on the organization and evolution of genes and genomes. The discovery and the characterization of repetitive elements such as transposable elements (TEs) can help to better understand the specific evolution of Triticeae species.

The current view of the TE composition in Triticeae genomes is based on a dominance of class 1 LTR retrotransposons and very few class 2 transposons. The increase and decrease of the genome size, the specific genome organization as well as DNA rearrangements were so far mainly explained by the activity of the LTR retrotransposons. Recently, numerous CACTA transposons were identified and annotated after the sequencing of 427 Kb in diploid and tetraploid wheat (Wicker et al. 2003b). Hybridization experiments based on the coding regions of one of the characterized sub-group called *Caspar* have shown the high redundancy of such elements in diploid wheat. Considering that only one quarter of all the annotated CACTA transposons seems to carry coding regions, the data suggest that these transposable elements contributed significantly to the genome size increase. Recently, a high copy number of a CACTA family was also identified in *Lolium perenne* (*Poeae*) and in other grass species, confirming our previous analysis (Langdon et al. 2003). The observation of the accumulation of class 2 transposons in cereals was supported by the observation of the high abundance of MITEs (Feschotte et al. 2002). MITEs are very small elements that are currently considered to be likely ancient class 2 transposon derivatives. Analysis of rice chromosome 4 has shown that MITEs constitute almost 50% of all the TEs identified (Feng et al. 2002). Similarly to LTR retrotransposons, transposons can now be considered as one of the major component of the Triticeae genomes.

Nevertheless, the mechanisms by which these elements are amplified remain speculative. It is known from studies on the transposition of bacterial transposons that mechanisms allowing the movement of transposons into new sites of the genome can also, depending on the timing of the

transposition and the DNA replication, increase the copy number of the transposon. This pathway called replicative transposition can create two copies of the transposons and is different from the conservative transposition mechanism that can be considered as a “cut and paste” mechanism. The molecular mechanisms preferentially supporting the replicative transposition rather than the conservative pathways in cereals genomes remain to be elucidated. However, the successful colonization of CACTA transposons in cereals indicates the high ability of the host genomes to tolerate their amplification, similarly to amplification of LTR retrotransposons.

In contrast to the class 1 retrotransposons that have been extensively studied, almost no data exist on the impacts of the high transposon redundancy on the organization, the evolution, DNA rearrangements and gene mutations that these elements can induce in the host genomes. It is known that class 2 transposons as well as MITEs exhibit a preferential location within euchromatic regions. In wheat, a substantial fraction of CACTA elements were identified from sequenced genomic segments of gene-rich regions (Wicker et al. 2003b). Furthermore, a CACTA transposon was found inserted within a coding region of the *Hv-β-GAL1* locus in barley (Guyot et al., manuscript in preparation). These data suggest that CACTA transposons are present in gene-rich regions of Triticeae and that they can influence the evolution of genes. Additional studies as well as the sequencing of large contiguous stretches of DNA in wheat will reveal the precise distribution of CACTA transposons among gene-rich and gene-poor regions and their impact on genome organization. We conclude that the amplification of CACTA transposon is one of the major mechanisms that participate to the genome evolution of Triticeae.

8.2 Large scale duplications in the ancestor of cereals contribute to the evolution of the Triticeae genome

The evolution of the wheat genome has also been studied by an *in silico* colinearity study between the rice genome and the distal part of the short arm of chromosome 1A in wheat. A total of 1.1 Mb of physical contig in which 638 kb were completely sequenced have been generated on the wheat chromosome 1A allowing us to investigate the micro-colinearity over large distances. While macro-colinearity has been found between wheat and rice, many micro-rearrangements such as deletions, inversions and gene movements indicated different mechanisms of the grass genome evolution. In addition, *in silico* analyses have revealed an intra-genomic colinearity in the rice genome suggesting an ancestral segmental duplication.

The extent of such duplications has been further studied using the availability of the pseudo-chromosomes sequences of rice. We found that the rice genome contains extensive chromosomal duplicated blocks accounting for 53% of the available sequences (Guyot and Keller 2004). Most of these duplications predate the divergence of most cereals (Goff et al. 2002; Vandepoele et al. 2003; Paterson et al. 2004) and must be conserved in the Triticeae lineage. The origin of the large duplicated blocks in rice remains controversial and two different mechanisms were proposed: polyploidy and aneuploidy. Ancestral polyploidy (or paleopolyploidy) is a well accepted theory probably because polyploidy is known to occur frequently in plants. Up to 80% of the angiosperms are estimated to be polyploids, ranging from the allotetraploidy of cotton, tetraploidy of maize, allohexaploidy of bread wheat, and up to 80 –ploidy in the *Sedum* genus (*Crassulaceae*). The complete sequencing of the small diploid genome of *Arabidopsis thaliana* has revealed the probability of three ancestral rounds of whole genome duplication (Simillion et al. 2002; Blanc et al. 2003). Recently, a combined strategy of comparative analysis between the rice structural genomic data and the genetic map of sorghum supported by a phylogenetic approach, indicated the presence of large duplicated blocks in rice. The extent and the dating of duplicated blocks suggested an ancestral polyploidization event occurring 70 MYA (Paterson et al. 2004) followed by a diploidization process (a process of return to disomy). In contrast, the observation of a non-uniform distribution of duplicated blocks and the fact that a large fraction of the genome was not found paired, suggested an aneuploid origin of rice and of most of the cereal genomes (Vandepoele et al. 2003) (Van de Peer Y., personal communication). However, the unequal distribution of blocks and the presence of non-duplicated regions can be the result of DNA rearrangements subsequent to the diploidization process. Further analysis and comparative analysis among grass species as well as a comparison of the duplications between rice and Arabidopsis genomes are required to better understand the origin of the rice duplications.

Whatever the origin (paleopolyploidy or aneuploidy), ancestral large duplications have an important impact on the evolution of the grass genomes. Chromosomal duplications have created a dramatic increase of the gene number in the ancestor of the cereal genomes. Such gene redundancy allowed copy of genes to escape the pressure of selection. These copies are free to be conserved by the genome and to evolve into new functions. Copy of genes can also be lost by the genome. In paleopolyploids such as in yeast and Arabidopsis, an extensive differential gene loss was reported (Dujon et al., 2004; Similion et al., 2002). Considering that duplications occurred

70 MYA (i.e. about 20 million years before the speciation of rice, maize and Triticeae from their common ancestor), these species have evolved independently during more than 40 million years. These divergence times after large scale duplications have probably contributed to generate variations of the conservation of the colinearity among cereals. We conclude that ancestral large-scale duplications have participated to the specific evolution of the Triticeae genome.

8.3 Segmental duplication and mechanisms of paralogous gene inactivation are involved in the evolution of the Triticeae genome

We analyzed the gene content of two paralogous and one orthologous segment at the barley and diploid wheat (*T. monococcum*) *β-galactosidase (β-GAL*) loci. Comparison of paralogous segments in barley indicated an ancestral origin through a segmental duplication followed by a movement of the duplicated segment from barley chromosome 5HL to chromosome 7HS. Comparative sequence analysis revealed mechanisms of paralogous gene inactivation such as deletion, mutation and transposon insertion in coding regions. Comparisons between barley paralogous regions and the rice genome revealed partial micro-colinearity interrupted by non-conserved coding regions, giving a mosaic organization of conserved sequences. These data indicate that gene inactivation and gene loss represent major factors in rapid evolutionary divergence of orthologous loci in Triticeae genomes (Chapter 7).

Gene inactivation was also found to contribute to the evolution of the high molecular weight (HMW)-glutenin orthologous regions of different wheat genomes (Gu et al. 2004). The characterization of more orthologous and paralogous regions from diploid and polyploid Triticeae species will further improve our understanding of these gene inactivation processes.

9 References

AGI. 2000. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408:796-815.

Ahn, S., J.A. Anderson, M.E. Sorrells, and T. S.D. 1993. Homoeologous relationships of rice, wheat and maize chromosomes. *Mol Gen Genet* 241: 483-490.

Akhunov, E.D., A.R. Akhunov, A.M. Linkiewicz, J. Dubcovsky, D. Hummel, G. Lazo, S. Chao, O.D. Anderson, J. David, L. Qi, and etal. 2003. Synteny perturbations between wheat homoeologous chromosomes caused by locus duplications and deletions correlate with recombination rates. *Proc Natl Acad Sci USA* 100: 10836–10841.

Altschul, S., T.L. Madden, A.A. Schaeffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25: 3389–3402.

Arumuganathan, K. and E.D. Earle. 1991. Nuclear DNA content of some important plants specie. *Plant Mol Biol Rep* 9: 208-218.

Belyayev, A., O. Raskina, and E. Nevo. 2001. Chromosomal distribution of reverse transcriptase-containing retroelements in two Triticeae species. *Chromosome Res.* 9: 129-136.

Bennett, M.D. and I.J. Leitch. 1995. Nuclear DNA amounts in angiosperms. *Ann Bot* 76: 113–176.

Bennett, M.D. and J.B. Smith. 1976. Nuclear dna amounts in angiosperms. *Philos Trans R Soc Lond B Biol Sci* 274: 227-274.

Bennetzen, J. 2000. Transposable element contributions to plant genome evolution. *Plant Mol Biol* 42: 251–269.

Bennetzen, J.L. and W. Ramakrishna. 2002. Numerous small rearrangements of gene content, order and orientation differentiate grass genomes. In *Plant Mol Biol*, pp. 821-827.

Blanc, G., A. Barakat, R. Guyot, R. Cooke, and M. Delseny. 2000. Extensive duplication and reshuffling in the Arabidopsis genome. In *Plant Cell*, pp. 1093-1101.

Blanc, G., K. Hokamp, and K.H. Wolfe. 2003. A recent polyploidy superimposed on older large-scale duplications in the Arabidopsis genome. *Genome Res* 13: 137-144.

Brooks, S.A., L. Huang, B.S. Gill, and J.P. Fellers. 2002. Analysis of 106 kb of contiguous DNA sequence from the D genome of wheat reveals high gene density and a complex arrangement of genes related to disease resistance. *Genome* 45: 963–972.

Brunner, S., B. Keller, and C. Feuillet. 2003. A large rearrangement involving genes and low copy DNA interrupts the microlinearity between rice and barley at the Rph7 locus. *Genetics* 164: 673-683.

Bureau, T. and S. Wessler. 1994a. Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. *Proc Natl Acad Sci U S A.* 91: 1411-1415.

Bureau, T. and S.R. Wessler. 1994b. Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Proc Natl Acad Sci USA* 9: 1411–1415.

Cannon, S.B., A. Kozik, B. Chan, R. Michelmore, and N.D. Young. 2003. DiagHunter and GenoPix2D: programs for genomic comparisons, large-scale homology discovery and visualization. *Genome Biol* 4: R68.

Cenci, A., N. Chantret, X. Kong, Y. Gu, O.D. Anderson, T. Fahima, A. Distelfeld, and J. Dubcovsky. 2003. Construction and characterization of a half million clone BAC library of durum wheat (Triticum turgidum ssp. durum). *Theor Appl Genet* 107: 931-939.

Chantret, N., A. Cenci, F. Sabot, O. Anderson, and J. Dubcovsky. 2004. Sequencing of the Triticum monococcum Hardness locus reveals good microcolinearity with rice. *Molecular Genetics and Genomics* 271: 377-386.

Chen, M., P. SanMiguel, and J.L. Bennetzen. 1998. Sequence organization and conservation in sh2/a1-homologous regions of sorghum and rice. *Genetics* 148: 435-43.

Chen, M., P. SanMiguel, A.C. deOliveira, S.S. Woo, H. Zhang, R.A. Wing, and J.L. Bennetzen. 1997. Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes. *Proc Natl Acad Sci U S A* 94: 3431-3435.

Chopra, S., V. Brendel, J. Zhang, J.D. Axtell, and T. Peterson. 1999. Molecular characterisation of a mutable pigmentation phenotype and isolation of the first active transposable element from Sorghum bicolor. *Proc Natl Acad Sci USA* 96: 15330–15335.

Cloix, C., S. Tutois, O. Mathieu, C. Cuvillier, M.C. Espagnol, G. Picard, and S. Tourmente. 2000. Analysis of 5S rDNA arrays in Arabidopsis thaliana:physical mapping and chromosome-specific polymorphisms. *Genome Res* 10: 679–690.

Cresse, A.D., S.H. Hulbert, W.E. Brown, J.R. Lucas, and J.L. Bennetzen. 1995. Mu1-related transposable elements of maize preferentially insert into low copy number DNA. *Genetics.* 140:315-24.

Devereux, J., P. Haeberli, and O. Smithies. 1984. A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res* 12: 387–395.

Devos, K., M.D. Atkinson, C.N. Chinoy, R. Harcourt, R.M.D. Koebner, C. Liu, P. Masojc, D.X. Xie, and M.D. Gale. 1993. Chromosomal rearrangements in the rye genome relative to that of wheat. *Theoretical Applied Genetics* 85: 784-792.

Devos, K.M., J.K. Brown, and J.L. Bennetzen. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* 12: 1075-1079.

Distelfeld, A., C. Uauy, S. Olmos, A.R. Schlatter, J. Dubcovsky, and T. Fahima. 2004. Microcolinearity between a 2-cM region encompassing the grain protein content locus Gpc-6B1 on wheat chromosome 6B and a 350-kb region on rice chromosome 2. *Funct Integr Genomics* 4: 59-66.

Dubcovsky, J., M. Luo, and J. Dvorak. 1995. Differentiation between homoeologous chromosomes 1A of wheat and 1Am of Triticum monococcum and its recognition by the wheat Ph1 locus. *Proc Natl Acad Sci U S A* 92: 6645-6649.

Dubcovsky, J., M. Luo, G. Zhong, R. Bransteitter, A. Desai, A. Kilian, A. Kleinhofs, and J. Dvorak. 1996. Genetic map of diploid wheat, T. monococcum L., and its comparison with maps of Hordeum vulgare L. *Genetics* 143: 983-999.

Dubcovsky, J., W. Ramakrishna, P.J. SanMiguel, C.S. Busso, L.L. Yan, B.A. Shiloff, and J.L. Bennetzen. 2001. Comparative sequence analysis of colinear barley and rice bacterial artificial chromosomes. *Plant Physiol* 125: 1342–1353.

Dunford, R.P., N. Kurata, D.A. Laurie, T.A. Money, Y. Minobe, and G. Moore. 1995. Conservation of fine-scale DNA marker order in the genomes of rice and the Triticeae. *Nucleic Acids Res* 23: 2724-2728.

Endo, T.R. and B.S. Gill. 1996. The deletion stocks of common wheat. *J Hered* 87: 295-307.

Feng, Q., Y. Zhang, P. Hao, S. Wang, G. Fu, Y. Huang, Y. Li, J. Zhu, Y. Liu, X. Hu, J. P., and etal. 2002. Sequence and analysis of rice chromosome 4. *Nature* 420: 316-320.

Feschotte, C., N. Jiang, and S.R. Wessler. 2002. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet* 3: 329-341.

Feschotte, C. and S.R. Wessler. 2002. Mariner-like transposases are widespread and diverse in flowering plants. *Proc Natl Acad Sci U S A* 99: 280-285.

Feuillet, C. and B. Keller. 1999. High gene density is conserved at syntenic loci of small and large grass genomes. *Proc Natl Acad Sci U S A* 96: 8265-8270.

-. 2002. Comparative genomics in the grass family: molecular characterization of grass genome structure and evolution. *Ann Bot Lond* 89: 3-10.

Feuillet, C., A. Penger, K. Gellner, A. Mast, and B. Keller. 2001. Molecular evolution of receptor-like kinase genes in hexaploid wheat: independent evolution of orthologs after polyploidization and mechanisms of local rearrangements at paralogous loci. *Plant Physiol* 125: 1304–1313.

Fu, H. and H.K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci U S A* 99: 9573-9578.

Gale, M.D. and K.M. Devos. 1998. Comparative genetics in the grasses. *Proc Natl Acad Sci U S A* 95: 1971-1974.

Gallego, F., C. Feuillet, M. Messmer, A. Penger, A. Graner, M. Yano, T. Sasaki, and B. Keller. 1998. Comparative mapping of the two wheat leaf rust resistance loci Lr1 and Lr10 in rice and barley. *Genome* 41: 328-336.

Gaut, B.S. 2002. Evolutionary dynamics of grass genomes. *New Phytol* 154: 15-28.

Gierl, A. and H. Saedler. 1989. Maize transposable elements. *Annu Rev Genet* 23: 71–85.

Gill, K.S., B.S. Gill, T.R. Endo, and E.V. Boyko. 1996a. Identification and high-density mapping of gene-rich regions in chromosome group 5 of wheat. *Genetics* 143: 1001-1012.

Gill, K.S., B.S. Gill, T.R. Endo, and T. Taylor. 1996b. Identification and high-density mapping of gene-rich regions in chromosome group 1 of wheat. *Genetics* 144: 1883-1891.

Goff, S.A., D. Ricke, T.H. Lan, G. Presting, R. Wang, and etal. 2002. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). *Science* 296: 92-100.

Graner, A., H. Siedler, A. Jahoor, R.G. Hermann, and G. Wenzel. 1990. Assessment of the degree and type of polymorphism in barley (Hordeum vulgare). *Theor Appl Genet* 80: 826-832.

Gu, Y.Q., D. Coleman-Derr, X. Kong, and O.D. Anderson. 2004. Rapid genome evolution revealed by comparative sequence analysis of orthologous regions from four triticeae genomes. *Plant Physiol* 135: 459-470.

Guyot, R. and B. Keller. 2004. Ancestral genome duplication in rice. *Genome* 47: 610-614.

Guyot, R., N. Yahiaoui, C. Feuillet, and B. Keller. 2004. In silico comparative analysis reveals a mosaic conservation of genes within a novel colinear region in wheat chromosome 1AS and rice chromosome 5S. *Funct. Int. Gen* 4: 47-58.

Han, B. and Y. Xue. 2003. Genome-wide intraspecific DNA-sequence variations in rice. *Curr Opin Plant Biol* 6: 134-138.

Harbert, N.F., RB. Thompson, RD. 1987. Identification of a transposon-like insertion in a *Glu-1* allele of wheat. *Mol. Gen. Genet.* 209:326-332.

Hoshino, A., Y. Johzuka-Hisatomi, and S. Iida. 2001. Gene duplication and mobile genetic elements in the morning glories. *Gene* 265: 1–10.

Huang, S., A. Sirikhachornkit, X. Su, J. Faris, B. Gill, R. Haselkorn, and P. Gornicki. 2002. Genes encoding plastid acetyl-CoA carboxylase and 3-phosphoglycerate kinase of the Triticum/Aegilops complex and the evolutionary history of polyploid wheat. *Proc Natl Acad Sci U S A* 99: 8133-8138.

Hudakova, S., W. Michalek, G.G. Presting, R. ten Hoopen, K. dos Santos, Z. Jasencakova, and I. Schubert. 2001. Sequence organization of barley centromeres. *Nucleic Acids Res* 29: 5029-5035.

Ilic, K., P.J. SanMiguel, and J.L. Bennetzen. 2003. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. *Proc Natl Acad Sci U S A* 100: 12265-12270.

Inagaki, Y., Y. Hitsatomi, T. Suzuki, K. Kasahara, and S. Iida. 1994. Isolation of a Suppressor-Mutator/Enhancer-like transposable element, Tpn1, from Japanese morning glory bearing variegated flowers. *Plant Cell* 6: 375–383.

Iwamoto, M. and K. Higo. 2003. Tourist C transposable elements are closely associated with genes expressed in flowers of rice (Oryza sativa). *Mol Genet Genomics* 268: 771-778.

Jiang, N., Z. Bao, X. Zhang, H. Hirochika, S.R. Eddy, S.R. McCouch, and S.R. Wessler. 2003. An active DNA transposon family in rice. *Nature* 421: 163-167.

Jurka, J. 2000. Repbase update: a database and an electronic journal of repetitive elements. *Trends Genet* 16:418-420.

Kalendar, R., C.M. Vicient, O. Peleg, K. Anamthawat-Jonsson, A. Bolshoy, and A.H. Schulman. 2004. Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. *Genetics* 166: 1437-1450.

Kashkush, K., M. Feldman, and A.A. Levy. 2003. Transcriptional activation of retrotransposons alters the expression of adjacent genes in wheat. *Nat Genet* 33: 102-106.

Keller, B. and C. Feuillet. 2000. Colinearity and gene density in grass genomes. *Trends Plant Sci* 5: 246-251.

Kellogg, E.A. 2001. Evolutionary history of the grasses. *Plant Physiol* 125: 1198-1205.

Kikuchi, K., K. Terauchi, M. Wada, and H.Y. Hirano. 2003a. The plant MITE mPing is mobilized in anther culture. *Nature* 421: 167-170.

Kikuchi, S., K. Satoh, T. Nagata, N. Kawagashira, K. Doi, N. Kishimoto, J. Yazaki, and etal. 2003b. Collection, mapping, and annotation of over 28,000 cDNA clones from japonica rice. *Science* 301:376-379.

Kishimoto, N., H. Higi, K. Abe, S. Arai, A. Saito, and K. Higo. 1994. Identification of the duplicated segments in rice chromosome 1 and 5 by linkage analysis of cDNA markers of known functions. *Theor Appl Genet* 88: 722-726.

Klein, P.E., R.R. Klein, J. Vrebalov, and J.E. Mullet. 2003. Sequence-based alignment of sorghum chromosome 3 and rice chromosome 1 reveals extensive conservation of gene order and one major chromosomal rearrangement. *Plant J* 34: 605-621.

Kong, X.Y., Y.Q. Gu, F.M. You, J. Dubcovsky, and O.D. Anderson. 2004. Dynamics of the evolution of orthologous and paralogous portions of a complex locus region in two genomes of allopolyploid wheat. *Plant Mol Biol* 54: 55-69.

Kumar, A. and J.L. Bennetzen. 1999. Plant retrotransposons. *Annu Rev Genet* 33: 479-532.

Kunzel, G.K., L. Meister, A. 2000. Cytologically integrated physical restriction fragment length polymorphism maps for the barley genome based on translocation breakpoints. *Genetics* 154:397-412.

Kurata, N., G. Moore, Y. Nagamura, T. Foote, M. Yano, Y. Minobe, and M. Gale. 1994. Conservation of genome structure between rice and wheat. *Bio-Technology* 12: 276-278.

La Rota, M. and M.E. Sorrells. 2004. Comparative DNA sequence analysis of mapped wheat ESTs reveals the complexity of genome relationships between rice and wheat. *Funct Integr Genomics* 4: 34-46.

Lagudah, E.D., J. Powell, W. 2001. Wheat Genomics. *Plant Physiol Biochem* 39:335-344.

Lander, E.S., P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln, and L. Newburg. 1987. MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics* 1: 174-181.

Langdon, T., G. Jenkins, R. Hasterok, R. Jones, and I. King. 2003. A high-copy-number CACTA family transposon in temperate grasses and cereals. *Genetics* 163: 1097-1108.

Leister, D., J. Kurth, D.A. Laurie, M. Yano, T. Sasaki, K. Devos, A. Graner, and P. Schulze-Lefert. 1998. Rapid reorganization of resistance gene homologues in cereal genomes. *Proc Natl Acad Sci U S A* 95: 370-375.

Lewin, B. 1997. Transposons. *In Genes VI. Oxford University Press, Inc, New York*: 563–595.

Li, C., P. Ni, M. Francki, A. Hunter, Y. Zhang, S. D., H. Li, A. Tarr, J. Wang, M. Cakir, J. Yu, M. Bellgard, R. Lance, and R. Appels. 2004. Genes controlling seed dormancy and pre-harvest sprouting in a rice-wheat-barley comparison. *Funct Integr Genomics* 4: 84-93.

Li, W. and B.S. Gill. 2002. The colinearity of the Sh2/A1 orthologous region in rice, sorghum and maize is interrupted and accompanied by genome expansion in the triticeae. *Genetics* 160: 1153-1162.

Lijavetzky, D., G. Muzzi, T. Wicker, B. Keller, R.A. Wing, and J. Dubcovsky. 1999. Construction and characterization of a bacterial artificial chromosome (BAC) library for the A genome of wheat. *Genome* 42: 1176–1182.

Lowe, T.M. and S.R. Eddy. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955-964.

Lynch, M. and J.S. Conery. 2000. The Evolutionary Fate and Consequences of Duplicate Genes. *Science* 290: 1151-115.

Ma, J., K.M. Devos, and J.L. Bennetzen. 2004. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res* 14: 860-869.

Ma, Z., S. Weining, P.J. Sharp, and C. Liu. 2000. Non-gridded library: a new approach for BAC (bacterial artificial chromosome) exploitation in hexaploid wheat (Triticum aestivum). *Nucleic Acids Res* 28: E106.

Manninen, I. and A.H. Schulman. 1993. BARE-1, a copia-like retroelement in barley (Hordeum vulgare L.). *Plant Mol Biol* 22: 829-846.

Miura, A., S. Yonebayashi, K. Watanabe, T. Toyama, H. Shimada, and T. Kakutani. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in Arabidopsis. *Nature* 411: 212–214.

Moullet, O., H.B. Zhang, and E.S. Lagudah. 1999. Construction and characterization of a large DNA insert library from the D genome of wheat. In *Theor Appl Genet*, pp. 305-313.

Murphy, G.J., H. Lucas, G. Moore, and R.B. Flavell. 1992. Sequence analysis of WIS-2-1A, a retrotransposon-like element from wheat. *Plant Mol Biol* 20: 991-995.

Nacken, W.K.F., R. Piotrowiak, H. Saedler, and H. Sommer. 1991. The transposable element TAM-1 of A. majus shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. *Mol Gen Genet* 28: 201–208.

Nacken, W.P., R. Saedler, H. Sommer, H. 1991. The transposable element Tam1 from Antirrhinum majus shows structural homology to the maize transposon En/Spm and has no sequence specificity of insertion. *Mol Gen Genet.* 228:201-208.

Nagaki, K., H. Tsujimoto, and T. Sasakuma. 1998. A novel repetitive sequence of sugar cane, SCEN family, locating on centromeric regions. *Chromosome Res* 6: 295–302.

Nakazaki, T., Y. Okumoto, A. Horibata, S. Yamahira, M.N. Teraishi, H., H. Inoue, and T. Tanisaka. 2003. Mobilization of a transposon in the rice genome. *Nature* 421: 170-172.

Neu, C., B. Keller, and C. Feuillet. 2003. Cytological and molecular analysis of the Hordeum vulgare-Puccinia triticina nonhost interaction. *Mol Plant Microbe Interact* 7: 626-633.

Ozeki, Y., E. Davies, and J. Takeda. 1997. Somatic variation during long term subculturing of plant cells caused by insertion of a transposable element in a phenylalanine ammonia-lyase (PAL) gene. *Mol Gen Genet* 254: 407–416.

Panstruga, R., R. Buschges, P. Piffanelli, and P. Schulze-Lefert. 1998. A contiguous 60 kb genomic stretch from barley reveals molecular evidence for gene islands in a monocot genome. *Nucleic Acids Res* 26: 1056-1062.

Paterson, A.H., J.E. Bowers, and B.A. Chapman. 2004. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. *Proc Natl Acad Sci U S A* 101: 9903-9908.

Peng, J., D.E. Richards, N.M. Hartley, G.P. Murphy, K.M. Devos, J.E. Flintham, J. Beales, L.J. Fish, A.J. Worland, F. Pelica, D. Sudhakar, P. Christou, J.W. Snape, M.D. Gale, and N.P. Harberd. 1999. 'Green revolution' genes encode mutant gibberellin response modulators. *Nature* 400: 256-261.

Pereira, A., H. Cuypers, A. Gierl, Z.S. Sommer, and H. Saedler. 1986. Molecular analysis of the En/Spm transposable element system of Zea mays. *EMBO J* 5: 835–841.

Presting, G.G., L. Malysheva, J. Fuchs, and I. Schubert. 1998. A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant J* 16: 721-728.

Qi, L., B. Echalier, B. Friebe, and B.S. Gill. 2003. Molecular characterization of a set of wheat deletion stocks for use in chromosome bin mapping of ESTs. *Funct Integr Genomics* 3: 39-55.

Raes, J., K. Vandepoele, C. Simillion, Y. Saeys, and Y. VandePeer. 2003. Investigating ancient duplication events in the Arabidopsis genome. *J Struct Funct Genomics* 3: 117-129.

Rahman, S., S. Abrahams, D. Abbott, Y. Mukai, M. Samuel, M. Morell, and R. Appels. 1997. A complex arrangement of genes at a starch branching enzyme I locus in the D-genome donor of wheat. *Genome* 40: 465-474.

Ramakrishna, W., J. Emberton, P. SanMiguel, M. Ogden, V. Llaca, J. Messing, and J.L. Bennetzen. 2002. Comparative sequence analysis of the sorghum Rph region and the maize Rp1 resistance gene complex. *Plant Physiol* 130: 1728-1738.

Rayburn, A.L. and B.S. Gill. 1986. Isolation of a G-genome specific sequence repeated DNA sequence from Aegilops squarrosa. *Plant Mol Biol Rep* 4: 102–109.

Rebatchouk, D. and J.O. Narita. 1997. Foldback transposable elements in plants. *Plant Mol Biol* 34: 831-815.

Rice, P., I. Longden, and A. Bleasby. 2000. EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics* 16: 276-277.

Rostoks, N., Y. Park, W. Ramakrishna, J. Ma, A. Druka, B.A. Shiloff, Z. Jiang, R. Brueggeman, D. Sandhu, K. Gill, and e. al. 2002. Genomic sequencing reveals gene content, genomic organization, and recombination relationships in barley. *Funct Integr Genomics* 2: 51–59.

Rubin, E., G. Lithwick, and A.A. Levy. 2001. Structure and evolution of the hAT transposon superfamily. *Genetics* 158: 949-957.

Sabot, F., R. Guyot, T. Wicker, N. Chantret, B. Laubin, B. Chalhoub, P. Leroy, P. Sourdille, and M. Bernard. 2004. Transposable Elements Content & Activities are Different among Wheat Large Genomic Sequences. *Manuscript in preparation*.

Salamini, F., H. Ozkan, A. Brandolini, R. Schafer-Pregl, and W. Martin. 2002. Genetics and geography of wild cereal domestication in the near east. *Nat Rev Genet* 30: 429-441.

Sandhu, D., J. Champoux, S. Bondareva, and K. Gill. 2001. Identification and physical localization of useful genes and markers to a major gene-rich region on wheat group 1S chromosomes. In *Genetics*.

Sandhu, D. and K.S. Gill. 2002a. Gene-containing regions of wheat and the other grass genomes. *Plant Physiol.* 128: 803-811.

-. 2002b. Structural and functional organization of the '1S08 gene-rich region' in the Triticeae. *Plant Mol Biol* 48: 791-804.

SanMiguel, P. and J.L. Bennetzen. 1998. Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. *Ann Bot* 82: 37–44.

SanMiguel, P., B.S. Gaut, A. Tikhonov, Y. Nakajima, and J.L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. *Nat Genet* 20: 43-45.

SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274: 765-768.

SanMiguel, P.J., W. RamaKrishna, J.L. Bennetzen, C. Busso, and J. Dubovsky. 2002. Transposable elements, genes and recombination in a 215-kb contig from wheat chromosome 5A(m). *Funct Integr Genomics* 2: 70–80.

Sasaki, T., T. Matsumoto, K. Yamamoto, K. Sakata, and a. et. 2002. The genome sequence and structure of rice chromosome 1. *Nature* 420: 312-316.

Schmidt, T. 1999. LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. *Plant Mol Biol* 40: 903-910.

Sears, E.R. 1966. Nullisomic-tetrasomic combinations in hexaploid wheat. In *Chromosome Manipulations and Plant Genetics* (ed. K.R. Lewis), pp. 29-45. Oliver and Boyd, Edinburgh.

Shirasu, K., A.H. Schulman, T. Lahaye, and P. Schulze-Lefert. 2000. A contiguous 66 kb barley DNA sequence provides evidence for reversible genome expansion. *Genome Res* 10: 908–915.

Simillion, C., K. Vandepoele, M. Van, M.C., M. Zabeau, and Y. VandePeer. 2002. The hidden duplication past of Arabidopsis thaliana. *Proc Natl Acad Sci U S A* 99: 13627-13632.

Smith, D. and R. Flavell. 1975. Characterisation of the wheat genome by renaturation kinetics. *Chromosoma* 50: 223–242.

Smith, D.B. and R.B. Flavell. 1974. The relatedness and evolution of repeated nucleotide sequences in the genomes of some Gramineae species. *Biochem Genet* 12: 243-256.

Snowden, K.C. and C.A. Napoli. 1998. PsI: a novel Spm-like transposable element from Petunia hybrida. *Plant J* 14: 43–54.

Song, R., V. Llaca, and J. Messing. 2002. Mosaic organization of orthologous sequences in grass genomes. In *Genome Res*, pp. 1549-1555.

Song, R. and J. Messing. 2003. Gene expression of a gene family in maize based on noncollinear haplotypes. *Proc Natl Acad Sci U S A* 100: 9055-9060.

Sonnhammer, E.L.L. and R. Durbin. 1995. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Reprinted from Gene Combis* 167: GC1–GC10.

Sorrells, M., R. La, M., C. Bermudez-Kandianis, R. Greene, and a. et. 2003a. Comparative DNA sequence analysis of wheat and rice genomes. In *Genome Res*.

Sorrells, M.E., R. La, M., C.E. Bermudez-Kandianis, R. Greene, and etal. 2003b. Comparative DNA sequence analysis of wheat and rice genomes. *Genome Res* 13: 1818-1827.

Stein, N., C. Feuillet, T. Wicker, E. Schlagenhauf, and B. Keller. 2000. Subgenome chromosome walking in wheat: a 450-kb physical contig in Triticum monococcum L. spans the Lr10 resistance locus in hexaploid wheat (Triticum aestivum L). *Proc Natl Acad Sci U S A* 97: 13436-13441.

Stephens, J.L., S.E. Brown, N. Lapitan, and D.L. Knudson. 2004. Physical mapping of barley genes using an ultrasensitive fluorescence in situ hybridization technique. *Genome* 47: 179-189.

Takahashi, S., Y. Inagaki, A. Hoshino, and S. Iida. 1999. Capture of a genomic HMG domain sequence by the En/Spm-related transposable element Tpn1 I the Japanese morning glory. *Mol Gen Genet* 261: 447–451.

Tarchini, R., P. Biddle, R. Wineland, S. Tingey, and A. Rafalski. 2000. The complete sequence of 340 kb of DNA around the rice Adh1-adh2 region reveals interrupted colinearity with maize chromosome 4. *Plant Cell* 12: 381-391.

Thompson, J.D., D.G. Higgins, and T.J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22: 4673–4680.

Tikhonov, A.P., P.J. SanMiguel, Y. Nakajima, N.M. Gorenstein, J.L. Bennetzen, and Z. Avramova. 1999. Colinearity and its exceptions in orthologous adh regions of maize and sorghum. *Proc Natl Acad Sci U S A* 96: 7409-7414.

Vandepoele, K., C. Simillion, and Y. VandePeer. 2003. Evidence that rice and other cereals are ancient aneuploids. *Plant Cell* 15: 2192-2202.

VanDeynze, A., J. Nelson, E. Yglesias, S. Harrington, D. Braga, S. McCouch, and M. Sorrells. 1995a. Comparative mapping in grasses. Wheat relationships. *Mol Gen Genet* 248: 744-754.

VanDeynze, A.E., J. Dubcovsky, K.S. Gill, J.C. Nelson, M.E. Sorrells, J. Dvorak, B.S. Gill, E.S. Lagudah, S.R. McCouch, and R. Appels. 1995b. Molecular-genetic maps of group 1 chromosomes of Triticeae species and their relations to chromosomes in rice and oat. *Genome* 38: 45-59.

Vicient, C., A. Suoniemi, K. Anamthawat-Jonsson, J. Tanskanen, A. Beharav, E. Nevo, and A. Schulman. 1999. Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. *Plant Cell* 11: 1769-1784.

Vicient, C.M., M.J. Jaaskelainen, R. Kalendar, and A.H. Schulman. 2001. Active retrotransposons are a common feature of grass genomes. *Plant Physiol* 125: 1283-1292.

Vision, T.J., D.G. Brown, and S.D. Tanksley. 2000. The origins of genomic duplications in Arabidopsis. *Science* 290: 2114-2117.

Wei, F., R.A. Wing, and R.P. Wise. 2002. Genome dynamics and evolution of the Mla powdery mildew resistance locus RT in barley. *Plant Cell* 14: 1903–1917.

Wendel, J.F. 2000. Genome evolution in polyploids. *Plant Mol Biol* 42: 225-249.

Wicker, T., R. Guyot, N. Yahiaoui, and B. Keller. 2003a. CACTA transposons in Triticeae. A diverse family of high-copy repetitive elements. *Plant Physiol.* 132:52-63.

Wicker, T., D.E. Matthews, and B. Keller. 2002. TREP: a database for Triticeae repetitive elements. *Trends Plant Sci* 7: 561–562.

Wicker, T., N. Stein, L. Albar, C. Feuillet, E. Schlagenhauf, and B. Keller. 2001. Analysis of a contiguous 211 kb sequence in diploid wheat (Triticum monococcum L.)reveals multiple mechanism of genome evolution. *Plant J* 26: 307–316.

Wicker, T., N. Yahiaoui, R. Guyot, E. Schlagenhauf, Z.D. Liu, J. Dubcovsky, and B. Keller. 2003b. Rapid genome divergence at orthologous LMW glutenin loci of the A and Am genomes of wheat. *Plant Cell* 15: 1186-1197.

Witte, C.P., Q.H. Le, T. Bureau, and A. Kumar. 2001. Terminal-repeat retrotransposons in miniature (TRIM) are involved in restructuring plant genomes. *Proc Natl Acad Sci U S A* 98: 13778-13783.

Wolfe, K.H. 2001. Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2: 333-341.

Wolfe, K.H., M. Gouy, Y.W. Yang, P.M. Sharp, and W.H. Li. 1989. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. In *Proc Natl Acad Sci U S A*, pp. 6201-6205.

Wolfe, K.H. and D.C. Shields. 1997. Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387: 708-713.

Wu, J., N. Kurata, H. Tanoue, T. Shimokawa, Y. Umehara, M. Yano, and T. Sasaki. 1998. Physical mapping of duplicated genomic regions of two chromosome ends in rice. *Genetics* 150: 1595-1603.

Yan, L., V. Echenique, C. Busso, P. SanMiguel, W. Ramakrishna, J.L. Bennetzen, S. Harrington, and J. Dubcovsky. 2002. Cereal genes similar to Snf2 define a new subfamily that includes human and mouse genes. *Mol Genet Genomics.* 268: 488-499.

Yan, L., A. Loukoianov, A. Blechl, G. Tranquilli, W. Ramakrishna, P. SanMiguel, J.L. Bennetzen, V. Echenique, and J. Dubcovsky. 2004. The wheat VRN2 gene is a flowering repressor down-regulated by vernalization. *Science* 303: 1640-1644.

Yan, L., A. Loukoianov, G. Tranquilli, M. Helguera, T. Fahima, and J. Dubcovsky. 2003. Positional cloning of the wheat vernalization gene VRN1. *Proc Natl Acad Sci U S A* 100: 6263-6368.

Yu, Y., J.P. Tomkins, R. Waugh, D.A. Frisch, D. Kudrna, A. Kleinhofs, R.S. Brueggeman, G.J. Muehlbauer, R.P. Wise, and R.A. Wing. 2000. A bacterial artificial chromosome library for barley (Hordeum vulgare L.) and the identification of clones containing putative resistance genes. *Theor. Appl. Genet.* 101:1093-1099.

Yuan, Q., S. Ouyang, J. Liu, B. Suh, F. Cheung, R. Sultana, D. Lee, J. Quackenbush, and C.R. Buell. 2003. The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res* 31: 229-333.

Zhang, Q., J. Arbuckle, and S.R. Wessler. 2000. Recent, extensive, and preferential insertion of members of the miniature inverted-repeat transposable element family Heartbreaker into genic regions. *Proc Natl Acad Sci USA* 97: 1160-1165.

Zhang, X., C. Feschotte, Q. Zhang, N. Jiang, W.B. Eggleston, and S.R. Wessler. 2001. P instability factor: an active maize transposon system associated with the amplification of Tourist-like MITEs and a new superfamily of transposases. *Proc Natl Acad Sci U S A* 98: 12572-12577.

10 Acknowledgments

I wish to thank my supervisor, Prof. Beat Keller, who offered me to do this PhD in his laboratory.
I am grateful to Prof Dr. U. Grossniklaus for taking charge of the co-referate.
I am grateful to Dr. Catherine Feuillet for her supervision and helpful discussions.
I am grateful to Dr. Thomas Wicker to introduce me during my diploma to the world of transposable elements.
I wish to thank Dr. C. Feuillet, Dr. B. Von Malek, Dr. E. Schlagenhauf, Dr. T. Wicker and Dr. N. Yahiaoui for their specific contribution on the scientific work presented in this thesis.
I am also grateful to Simon Krattinger and Barbara Schellenberg for the german translation of the summary.
Finally, I am grateful to all members and former members of the Beat Keller ‘s laboratory for providing me continuous help.

11 Curriculum Vitae

Guyot

Romain

20/06/1971

French

Lycée Dr. Lacroix, Narbonne (Aude) France, Baccalauréat, 1990

University of Toulouse France, Biology degree in physiology and cell biology, 1996

University of Zürich Switzerland, Diploma in Plant Biology : dipl. Bot, Institute of Plant Biology, Department of Plant Molecular Biology, Prof. Dr. Beat Keller., October 2002.

Title of the diploma thesis : "Comparative analysis suggests rapid evolution of orthologous loci in diploid and tetraploid wheat"

University of Zürich Switzerland, Ph.D. studies Institute of Plant Biology, Department of Plant Molecular Biology, Prof. Dr. Beat Keller.

Title of the Ph.D. thesis: "Mechanisms of Triticeae Genome Evolution"

At the University of Zurich since 2001

Publications during diploma and doctorate studies :

Whole-genome comparison of leucine-rich repeat extensins in Arabidopsis and rice. A conserved family of cell wall proteins form a vegetative and a reproductive clade. (2003) Baumberger N, Doesseger B, Guyot R, Diet A, Parsons RL, Clark MA, Simmons MP, Bedinger P, Goff SA, Ringli C, Keller B. *Plant Physiol*.131:1313-1326

Dynamic genome evolution at orthologous LMW glutenin loci of the A and A^m genomes of wheat. (2003) Wicker T, Yahiaoui N, Guyot R, Dubcovsky J, Schlagenhauf E, Liu ZD and Keller B**.** *Plant Cell*. 15:1186-1197

CACTA transposon in Triticeae – a diverse family of high-copy repetitive elements. (2003) Wicker T, Guyot R, Yahiaoui N and Keller B. *Plant Physiol*.132:52-63

A new structural element in the protoxylem of seed plants contains glycine-rich proteins. (2004) Ryser U., Schorderer M., Guyot R. and Keller B. *J Cell Sci*. 117:1179-1190

In silico comparative analysis reveals a mosaic conservation of genes colinear region in wheat chromosome 1AS and rice chromosome 5S. (2004) Guyot R, Yahiaoui N, Feuillet C. and Keller B. *Funct Integr Genomics*. 2004 4:47-58

Ancestral genome duplication in rice. (2004) Guyot R. and Keller B. *Genome*. 47:610-614